UNIVERSIDADE FEDERAL DE MINAS GERAIS
Faculdade de Filosofia e Ciências Humanas
Departamento de Filosofia

Carlos Henrique Barth

REPRESENTATIONAL COGNITIVE PLURALISM: towards a cognitive science of relevance-sensitivity

Belo Horizonte
2024

Carlos Henrique Barth

REPRESENTATIONAL COGNITIVE PLURALISM: towards a cognitive science of relevance-sensitivity

Tese apresentada ao Programa de Pós-graduação em Filosofia da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Doutor em Filosofia.

Orientador(a): Ernesto Perini Frizzera da Mota Santos

Belo Horizonte

2024

UNIVERSIDADE FEDERAL DE MINAS GERAIS
FACULDADE DE FILOSOFIA E CIÊNCIAS HUMANAS
PROGRAMA DE PÓS-GRADUAÇÃO EM FILOSOFIA

FOLHA DE APROVAÇÃO

Pluralismo cognitivo representacional: rumo a uma ciência cognitiva da sensibilidade à relevância

CARLOS HENRIQUE BARTH

Tese submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em FILOSOFIA, como requisito para obtenção do grau de Doutor em FILOSOFIA, área de concentração FILOSOFIA, linha de pesquisa Lógica, Ciência, Mente e Linguagem.

Aprovada em 23 de fevereiro de 2024, pela banca constituída pelos membros:

Prof. Ernesto Perini Frizzera da Mota Santos - Orientador (UFMG)

Prof. André Joffily Abath (UFMG)

Prof. Marco Aurélio Sousa Alves (UFSJ)

Prof. Felipe Nogueira de Carvalho (UFLA)

Profa. Maria Eunice Quilici Gonzalez (UNESP)

Belo Horizonte, 23 de fevereiro de 2024.

Documento assinado eletronicamente por **Felipe Nogueira de Carvalho**, **Usuário Externo**, em 26/02/2024, às 10:04, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Maria Eunice Quilici Gonzalez**, **Usuária Externa**, em 26/02/2024, às 14:14, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Ernesto Perini Frizzera da Mota Santos**, **Professor do Magistério Superior**, em 26/02/2024, às 14:24, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Marco Aurélio Sousa Alves**, **Usuário Externo**, em 26/02/2024, às 15:34, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

To my wife Rochelle, the most relevant part of my life.

AGRADECIMENTOS

Writing this became more painful than it should. There's just too many who can't hear how grateful I am for their love and support. My father, endless source of inspiration, who always wanted to be closer and closer to his loved ones. In my case, even by learning philosophy and trying to understand what the hell his son got into. (Your book is here with me now, with the bookmark still on the same page where you left it). My father-in-law, who was so enthusiastic that he arrived before me at my master's defense. Moreover, he wanted to make some f̈riendly remarksẗo one of the professors on the panel after hearing him make what he considered ünnecessary appointmentsd̈uring the defense (Thanks Rochelle for stopping him). I really miss that sparkle in his eyes. Stories like these make me feel a bit better, but they don't erase the wish that both of them could be here celebrating with me.

With all my heart, I want to thank my love Rochelle for believing, understanding and being there. You made this work, as well as the dreams surrounding it, possible. You deserve the best of me, always. Going on, I want to say a big T̈hank youẗo my mother and my sister, always rooting for me. To the new star in my life, Igor (who just can't stop making me proud), for worrying about what's on my head. To my advisor Ernesto Perini, for the helpful discussions, support, trust and the freedom to pursue my own path. Everything that's right about this work is also thanks to him (as for what's wrong, the fault is all mine). To the philosophy professors at UFMG, who contributed so much to my education, specially André Abath, for helpful discussions, encouragement and support. To Marco Aurélio Sousa Alves and Eros Carvalho, who provided valuable feedback after reading a previous version of this work. To the friends from philosophy, who shared the same academic journey, you have made this endeavor much more stimulating and enjoyable, specially Verônica Campos e Daniel Debarry. I miss the coffees, the conversations and the happy hours. To long term friends, Jean Francener, Rafaeli Ianegitz, Tiago Scheuer and Mariana Winck (I promise I'll never try to explain what this work is about).

Finally, I would like to thank the professors on the panel. André Abath, Felipe Carvalho, Maria Eunice Quilici Gonzales and Marco Aurélio Sousa Alves. Everyone was so kind and understanding that they agreed to read this material during their vacation time. Thank you so much! I would also like to thank all the members of the philosophy department, for all their invaluable administrative support in the course of this long journey. To the CAPES, for their financial support. And last, but not least, to Alexandra Sanovna Elbakyan. The Brazilian scientific and philosophical endeavors owe her more than they're ready to admit.

Beware of bugs in the above code; I have only proved it correct, not tried it. (Donald Knuth)

RESUMO

A tese busca contribuir para a explicação de duas capacidades cognitivas fundamentais da inteligência humana denominadas senso comum e holismo situacional. Nas ciências cognitivas, a tarefa de explicar essas capacidades enfrenta um desafio fundacional: como é possível à cognição humana distinguir o que é ou não relevante em um número indefinidamente multiplicável de contextos? O desafio se caracteriza por uma circularidade: potenciais soluções acabam pressupondo aquilo que pretendiam explicar, i.e. a sensibilidade ao que é contextualmente relevante. O argumento desenvolvido é uma tentativa de escapar dessa circularidade. Ele tem por base uma análise bidimensional do problema. A primeira dimensão é representacional e abarca como o conhecimento cognitivo sobre o que é relevante pode ser armazenado. Ela se associa ao ”frame problem” da inteligência artificial. Na tese argumenta-se que a dimensão representacional do problema pode ser neutralizada pelo abandono de esquemas representacionais linguísticos em virtude de uma pluralidade de esquemas representacionais estruturais. Estes permitem que estados representacionais possam exercer um papel explicativo claro e não trivial, e ao mesmo tempo permitem um tipo de produtividade representacional que evita o ”frame problem”. A segunda dimensão é inferencial. Ela abarca a sensibilidade à relevância que caracteriza nossa produtividade inferencial, i.e. nossa capacidade de inferir o que é relevante em um número indefinidamente multiplicável de contextos, incluindo aquelas que nunca encontramos antes. A esse respeito, argumenta-se que a sensibilidade à relevância não pode ser nem inata, pois mecanismos inatos não podem evitar a circularidade mencionada, nem aprendida através de mecanismos inatos, pois processos de aprendizado tomam a sensibilidade à relevância como pressuposto. A alternativa sugerida é que a sensibilidade à relevância é um *gadget* cognitivo. Um tipo de mecanismo mental que é herdado não geneticamente, mas culturalmente. Essa possibilidade é viabilizada a partir do trabalho de Cecilia Heyes, que descreve como vieses de desenvolvimento culturalmente sedimentados podem informar a constituição de mecanismos cognitivos. Assim, fatores culturais participam da cognição não apenas na forma de informação disponível para processamento, mas também como elementos que enviesam esse processamento na direção daquilo que é culturalmente estabelecido como relevante em cada contexto.

Palavras-chave: frame problem; problema da relevância; contexto; ciência cognitiva; representações mentais; mecanismos mentais.

ABSTRACT

This work aims to contribute to the explanation of cognitive capacities that are essential to human intelligence: commonsense and situation holism. The attempt to explain them within the field of cognitive sciences raises a foundational challenge. How can human cognition distinguish what's relevant and what's not in an open-ended set of contexts? The challenge is characterized by a circularity. Potential solutions end up relying on the very capacity that they should be explaining, i.e. the sensitivity to what's contextually relevant. The argument here developed is an attempt to get out of this circularity. It is grounded on a bi-dimensional analysis of the problem. The first one is representational and is about how cognitive knowledge on what's relevant can be stored. This dimension is related to the frame problem from artificial intelligence. I claim that this representational aspect of the issue can be neutralized by abandoning language-like representational schemes and sticking to a plurality of structural ones. They allow representational states to play a clear and non-trivial explanatory role while enabling a kind of representational productivity that is not bothered by the frame problem. The second dimension is the inferential one. It encompasses our relevance-sensitive inferential productivity, i.e. our capacity to efficiently infer what's relevant in indefinitely many contexts, including ones we've never faced before. I argue that such relevance sensitivity cannot be neither innate, for innate mechanisms result in the aforementioned circularity, nor learned through innate mechanisms, for learning processes themselves rely on relevance sensitivity. The suggested alternative is that relevance sensitivity is a cognitive gadget. A kind of mental mechanism that is culturally, rather than genetically, inherited. This possibility is rendered plausible within Cecilia Heyes' ”cognitive gadgets” framework. She shows how developmental tweaks and biases culturally established can participate in the constitution of cognitive mechanisms. Thus, cultural elements can participate in human cognition not only as information sources, but also in the form of tweaks that bias the processing of that information towards what's culturally established as contextually relevant.

Keywords: frame problem; relevance problem; context; cognitive science; mental representations; mental mechanisms.

LIST OF FIGURES

Figure 1 – A-B-C spatial relationships. 107
Figure 2 – A red rectangle. 117
Figure 3 – Palette-C. A scheme for representing simple cartoons. 126
Figure 4 – Embedding 2D shapes in 3D schemes 127
Figure 5 – Embedding 3D shapes in 2D schemes 128
Figure 6 – Structural similarity between two nodes. 134
Figure 7 – Theorem 1 from Galileo 137
Figure 8 – The smallest set of maximally independent basis functions for constructing natural images. 169
Figure 9 – The Palette-C scheme, again. 170
Figure 10 – The causal structure employed in the experiment. 174
Figure 11 – Churchland's prototypical partitions. 207
Figure 12 – Structural contents in partitions of activation spaces. 208
Figure 13 – What's relevant about daddy? 231
Figure 14 – A normal checkerboard. 232
Figure 15 – A mutilated checkerboard. 233
Figure 16 – Imitation in children. 246

LIST OF TABLES

Table 1 – Toy-like cognitive system. 35
Table 2 – How the human body and brain “represents” weight — Adapted from Povinelli (2012). 193
Table 3 – Input-output pairs in training set- from Clark; Thornton (1997). 234
Table 4 – Derived inputs - from Clark; Thornton (1997). 234

LISTA DE ABREVIATURAS E SIGLAS

AI	Artificial intelligence
CRC	Continuous Reciprocal Causation
FP	Frame problem
HPC	Hard problem of content
PHP	Penn, Holyak and Povinelli (proponents of the relational redescription hypothesis)
PSS	Physical symbol system
RCP	Representational cognitive pluralism
RP	Relevance problem
SPAC	Special-purpose adaptive couplings
SAH	Symbol approximation hypothesis

CONTENTS

INTRODUCTION . . . 14

1 THE RELEVANCE OF RELEVANCE . . . 22

1.1 What intelligence amounts to . . . 22

1.1.1 Commonsense holism . . . 24

1.1.2 Situation holism . . . 28

1.2 The non-saturable character of human contexts . . . 30

1.3 The role of context in cognitive explanations . . . 40

1.4 How context sensitivity raises problems about relevance . . . 45

1.5 Why relevance problems are so hard . . . 48

1.5.1 On Fodor's pessimism . . . 49

1.5.1.1 Can we dodge Fodor through heuristics? . . . 53

1.5.1.2 Can we dodge Fodor through massive modularity? . . . 57

1.5.1.3 Taking stock . . . 62

1.5.2 On simplifying things by bringing Heidegger . . . 62

1.5.2.1 The Wheeler-Dreyfus debate on the role of representational contents . . . 64

1.5.2.2 The Wheeler-Dreyfus-Rietveld debate on relevance sensitivity . . . 68

1.5.2.3 Taking non-representational stock . . . 75

1.5.3 What about this nice framework here? . . . 77

1.6 Mind the gap . . . 80

1.7 Representational and inferential productivity . . . 82

1.8 Paragraph-by-paragraph summary . . . 84

2 REPRESENTATIONAL PRODUCTIVITY . . . 92

2.1 The challenge of representational productivity . . . 92

2.1.1 For the want of a frame . . . 95

2.1.2 What is the frame problem about? . . . 100

2.2 Structural representations and the frame problem . . . 104

2.2.1 What structural representations buy us . . . 106

2.2.1.1 Representing change . . . 108

2.2.1.2 Intrinsic contents in structural representations . . . 110

2.2.2 The explanatory role of architectural constraints . . . 111

2.2.3 How many schemes would you say there are? . . . 114

2.2.4 Representing vs. Encoding . . . 116

2.2.5 Against monism: the incommensurability thesis . . . 119

2.2.6 Individuating representational schemes through contents . . . 125

2.2.7 A final diagnostic for FP . . . 129

2.3 Representational pluralism . . . 131

2.3.1 Representational redescription . . . 132

2.3.1.1 Complex token articulation as food for thought . . . 135

2.3.1.2 Playing with food: encoding . . . 137

2.3.1.3 Playing with food: task-embedding . . . 139

2.3.1.4 Still playing with food: analogical reasoning . . . 140

2.4 Explosion-free representational productivity . . . 141

2.5 Getting ready for what comes next . . . 144

2.6 Paragraph-by-paragraph summary . . . 145

3 THE APPLICABILITY OF REPRESENTATIONAL PLURALISM . . . 156

3.1 Battle of frameworks . . . 156

3.2 Representations and targets . . . 160

3.3 What are mental representations? . . . 162

3.3.1 Do we actually exploit structural correspondences? . . . 167

3.3.1.1 First example: natural images . . . 168

3.3.1.2 Second example: rats, mazes and food . . . 172

3.3.1.3 Third example: causal structures . . . 174

3.3.2 How to tell whether one is exploiting representations? . . . 175

3.3.3 The role of representational contents in the assembly of attitude contents . 180

3.4 What representations buy us . . . 184

3.4.1 What most representationalists got wrong . . . 188

3.5 Apes, humans and representational redescription . . . 192

3.5.1 What's wrong with the symbol approximation hypothesis . . . 196

3.5.2 The relational redescription hypothesis . . . 200

3.5.2.1 Domain bootstrapping . . . 202

3.5.2.2 Similarity judgments within emergent domains . . . 206

3.6 What now? . . . 209

3.7 Paragraph-by-paragraph summary . . . 210

4 INFERENTIAL PRODUCTIVITY . . . 221

4.1 Where are we at? . . . 221

4.2 Effective connectivity . . . 225

4.3 Can we learn what's relevant? . . . 229

4.3.1 Option one: general-purpose learning . . . 236

4.3.2 Option two: meta-learning . . . 237

4.3.3 Option three: innate tweaks . . . 240

4.3.4 Option four: incremental learning . . . 240

4.4 The cognitive gadgets framework . . . 242

4.5 Relevance sensitivity as a cognitive gadget . . . 249
4.5.1 Storing relevance-sensitive knowledge about the world . . . 252
4.5.2 Relevance-sensitive influence over cognitive processing . . . 254
4.5.3 Relevance-sensitive context extrapolation . . . 256
4.5.4 Relevance-sensitive context switching . . . 260
4.6 Closing remarks . . . 263
4.7 Paragraph-by-paragraph summary . . . 265

CONCLUSION . . . 274

BIBLIOGRAPHY . . . 281

INTRODUCTION

> Nothing is more admirable, than the readiness, with which the imagination suggests its ideas, and presents them at the very instant, in which they become necessary or useful. The fancy runs from one end of the universe to the other in collecting those ideas, which belong to any subject. One would think the whole intellectual world of ideas was at once subjected to our view, and that we did nothing but pick out such as were most proper for our purpose. There may not, however, be any present, beside those very ideas, that are thus collected by a kind of magical faculty in the soul, which, tho' it be always most perfect in the greatest geniuses, and is properly what we call a genius, is however inexplicable by the utmost efforts of human understanding. (Hume, 1965, p. 24)

Humans can zero in on what's circumstantially relevant to their goals and expectations. They can do that for indefinitely many goals within an open-ended set of distinct circumstances. How is that possible? Whenever this question is posed, it is not unusual to hear replies like: "Impressive, but what's so mysterious about it?" Indeed, showing why it is hard to explain this capacity requires some work.

One of the reasons behind the required effort is that we're used to seeing relevance as a solution rather than as a problem. In most philosophical and scientific endeavors, relevance sensitivity plays the role of an explanatory resource. For instance, communication is possible because, among other things, we can stick to the relevant portion of our common background of shared knowledge. As communication develops, we're expected to continuously review and constrain this set. Otherwise, the result can be hardly considered communication. It would be impossible to communicate with somebody that, despite being able to make perfectly valid inferences, simply can't stick to the relevant ones. Similarly, whenever we need to justify a conclusion, we can explain the presence of inferential steps by appeal to their relevance for the task at hand. One can say, for instance, that a given inference is there because it helps to select between two plausible possibilities in the course of an investigation. In this sense, relevance is a tool available for accounts of rationality, language, epistemology, etc.

But relevance sensitivity is also a problem. It emerges when trying to account for the cognitive capacities underpinning its aforementioned explanatory roles. More specifically, it comes to the fore whenever one is trying to fully understand how those mechanisms work by modeling them from the scratch. This is so because only in this kind of situation one is required to worry about everything that can bear on the exercise of some cognitive capacity or performance. As an instance of this point, consider a situation in which we're instructing someone on how to cook a certain dish. We expect them to be able to understand an instruction such as "stir until smooth". This means, among other things, that we don't expect to have to specify that the verb "stir" concerns the dish rather than the cutlery (which could be stirred out of the pan) or the body of the person receiving the instruction. We don't expect this because we rely on the learner's good judgment. But if we're trying to model the mechanics of the capacity to understand that instruction, there's no such thing to rely on. We have no choice but to address directly questions like: given everything that is known, how to select just the portions that can bear on the interpretation of that utterance? That is, how to select the relevant portions? Either the model somehow prevents inappropriate interpretations of the utterance, or it will fail as an explanation of why a certain understanding was obtained. The problem is that what's relevant can vary drastically from circumstance to circumstance, and the model must account for every single one of them.

Once we find ourselves in this kind of situation, we're forced to mitigate the model's explanatory reach. This means that, rather than a model accounting for (say) the capacity to understand English, the model will only account for the understanding of a single sentence in a very peculiar circumstance (or perhaps a small set of similar sentences in a very restricted set of circumstances). That is, we avoid the problem of handling the huge amount of possible circumstances by artificially (and somewhat arbitrarily) narrowing them down. Although this description makes it sound as if this type of extremely restricted model is useless, this is not at all the case. A lot of artificial intelligence (AI) models such as those powering chess engines are instances of this kind. The model concentrates on what's relevant for (say) playing chess and ignores everything else. The point is that, in those models, the ability to stick to what's relevant is completely put aside, for the applicability of the model relies on the relevance sensitivity of those using it. We are the one's responsible to apply the chess model only when the ability at stake is that of playing chess rather than, say, to cook.

How can we escape from this predicament? An obvious yet implausible possibility is to rely on "brute force" and simply start figuring out what's appropriate to every circumstance. We do that by throwing assumptions into the model in the form of conditional rules ("if the situation is such, do that"). Thus, the model can predict different outcomes for different circumstances, as long as they're covered by its set of rules. The problem now is to decide when to stop. The number of required rules can quickly escalate and get out of control, given the huge amount of possible circumstances. We need to stop somewhere, and the arbitrariness may be unavoidable ("Enough! This is going to be a model of how one understand a certain subset of possible English sentences involved in learning how to cook with certain instruments, certain ingredients and certain conditions of pressure and temperature."). But whenever we stop, all we have is again a model with a mitigated explanatory reach. It still relies on our ability, as designers or users of the model, to tell the kinds of circumstances in which the model can be applied from those in which it can't. This is not hard to see because, even if the model covers a relatively large number of circumstances, it's relatively easy to come up with unaccounted ones.

Roughly, this is the general form in which relevance issues present themselves when modeling cognitive abilities. All that's needed is a complex dispositional capacity — not necessarily linguistic — whose output may change in virtue of an open-ended set of circumstantial elements. While modeling, either we stop somewhere after covering for a relatively small subset of the possible permutations, or we watch them grow without control as a combinatorial explosion takes place. In both cases, the target capacity is left unexplained. But of course even after this first contact with the problem, there is still room for complaints: OK, but what's really so difficult about getting away from it? Perhaps our cognition is just a collection of domain-limited or special-purpose capacities? Under this optimistic view, as we create more and more limited-reach models of specific abilities, we're closer and closer to the full picture. There's no reason for despair, is there?

In my view, this is just one of the many possible ways of confusing a reformulation of the problem with a potential solution (or dissolution). The core motivation behind this particular suggestion is the thesis that cognition can be somehow analyzed into mechanisms that can be fully understood in isolation. But this is a rather strong requirement. It takes its toll later, when the need to articulate such special-purpose mechanisms in indefinitely many ways shows itself. The question becomes: what is the articulation of special-purpose mechanisms that captures what's relevant in this particular case? This is unavoidable, for we need to account for the productive way in which we can reuse what we know in an open-ended set of circumstances. Otherwise, we would need a special-purpose mechanism for each circumstance, which is unrealistic at best.

It is not hard to see that learning mechanisms suffer from the same issue. Learning without any kind of supervision amounts to a task like this: given everything I know, what should I infer now? Well, presumably one must infer what is relevant for one's goals. But that amounts to a new formulation of the same problem: given everything I know, which among the possible inferences I can do should be regarded as relevant for my goal? Again, either we drastically (and artificially) narrow down the set of possibilities and concentrate on the inferences leading to the domain that we (as theorists) are interested in modeling (e.g. learning about chess), or we fall victims of the combinatorial explosion that follows. Notice also that having a learning supervisor is not a direct answer to this problem. Whenever a mechanism is trying to acquire or improve some skill by following something else, it must be capable of handling the instructions correctly. But instruction sets, feedback and encouragement also rely on one's "good judgment", i.e. on one's capacity to figure out what's relevant in order to properly follow them (it was no accident that the cooking example above took place in a learning scenario).

The moral of stories like this is the following: when it comes to relevance sensitivity, it's simply too easy to confuse a reformulation of the problem with a solution or dissolution. If we look back at the history of how problems of relevance have manifested themselves in the cognitive sciences and AI, we'll find several cases of this kind, and some of them will be discussed in the course of this work. But looking back for its traces in early cognitive science work also shows that, most of the time, it was disregarded as a problem for the future or for somebody else. Most researchers were simply not worried about it. Part of the explanation, I suspect, lies in the aforementioned thesis that cognitive intelligence can be broken down and understood in the form of independent mechanisms. The problem, to be clear, is not that it is possibly false, but rather taking it for granted. Though that makes intuitive sense, the requisites for a model that attempts to explain a capacity in isolation are rather different from those that must make room for its further integration with other mechanisms. In other words, Occamn's razor is not necessarily a friend in this case. The simplest model for a capacity taken in isolation may leave out features that are required for its integration with the encompassing system. But that worry was usually put aside, and relevance issues was broadly regarded as a minor issue for when the computers in which we run our models become powerful enough for us to consider more realistic and integrative scenarios.

However, relevance issues emerged much earlier than predicted, and they did that through a quite curious route. In the 1960s, Patrick Hayes and John McCarthy presented the first description of the *frame problem* as a practical (yet deep) problem they were facing in their logicist approach to AI. As Hayes argues (probably until today) and as I'll argue in chapter 2, the frame problem must be distinguished from relevance issues: it is not really about relevance, but rather about modelling change. But what matters for us is that it was taken to be a problem about relevance by philosophers like Dennett and Fodor, and that escalated quickly. That's why even today, much discussion uses "frame problem" to denote relevance issues. It was this confusion, not more powerful computers or increasingly comprehensive models, that brought relevance issues to the fore. Perhaps even more curiously, with today's available computational power and gigantic linguistic models such as those from the GPT family, some still have a hard time in seeing the relevance of relevance issues. But now, rather than a problem "for the future", it is sometimes taken as a harmless ghost from the past. A reminiscence of when we still had to generate computer models of cognitive capacities "by hand" instead of using powerful machine learning techniques. Unfortunately, like any other learning, machine learning also must answer the crucial question: given everything I know what should I infer now? And even up to this day, that's only feasible within artificially circumscribed domains, such as that of recognizing faces or playing a game.[1]

The history of relevance problems is thus full of different diagnoses of what is to blame and prescriptions of what should be done about it. Some take it to be a problem with our modeling tools: they simply can't express what's needed. Thus, to change the modelling clay being used might be enough to buy us what's needed. Others take it to be a metaphysical issue: we must partition the world in the right set of joints. The details will be properly discussed and need not bother us now. What matters is to know that much of the dialectics employed in the chapters to come was adopted with this in mind. As we tell the story of *prima facie* good answers later revealed as naive, the hope is that it helps the reader in realizing how deep the issue goes. That's why considerable effort was made to carefully discuss matters that are sometimes conflated in the literature.

But this work does not only comprise an interpretation of the failed attempts. It tries to make a positive contribution as well (and of course risks becoming just another chapter in the

[1] Aren't current large language models (LLMs) — such as those using the GPT family — evidence to the contrary? They do seem to entertain proper conversations in a quite large number of circumstances, don't they? Yes, they seem to do that, but not in a way that throws light on how we handle relevance. LLMs capture a single domain, which is our usage pattern of syntactic elements in natural language. They capture what we do in a lot of distinct circumstances, but they don't capture why or how. What we expect from it is similar to what we would expect from a tidal frequency model. It can be useful for making predictions about tides, but it doesn't shed light on the mechanisms that cause tidal dynamics to exist and have the shape they do.

history of failed attempts). I locate it at the philosophy of cognitive sciences, a place that is much more like philosophy of science than it is philosophy of mind. If there is any hope for cognitive science's general project of providing grounds for epistemically assessing biological or computational states and processes, then we must be able to explain how capacities that rely on relevance sensitivity are possible. More precisely, we must be able to formulate empirical hypotheses about how it is done. Thus, we must make sure that we're not being constrained nor misled by foundational posits from our favorite cognitive framework or research tradition. In other words, as Chemero (2009) warned us, we need to make sure that we're not distorting or ignoring empirical possibilities through non-empirical means.

As an instance of how this was handled, consider the use of representations in accounts of cognitive capacities. I've tried to develop the theses I'm presenting in a predominantly representational framework. The goal is not to claim that representations are necessary to handle relevance issues, though. It is to show that problems of relevance do not arise directly from the use of representations (even though some kinds of representation can put a break on the representationalist's hope, as we'll see). Consequently, even if it turns out to be empirically true that cognition does rely on representational contents in some cases, that need not worry us any more than it does friends of broadly non-representational frameworks, such as enactivism and ecological psychology.

With this in mind, I can now present the two core claims of this work:

(1) relevance sensitivity is a *cognitive gadget*; and
(2) a representational account of this gadget can be forged by using frameworks compatible with *representational cognitive pluralism.*

Of course at this point that doesn't tell us much. This is why in what follows I provide the core structure of the argument for these claims, as they are presented in the forthcoming chapters. Sometimes the discussions need to go deep in quite specific matters, and there's always the risk to lose track of the big picture. Hopefully, the following synthesis can be of help with that.

The first chapter concentrates on characterizing what will be called the *relevance problem* (RP). At first, it tries to do so in a rather abstract way. This is important, for a vital point of this chapter is that RP is not a feature of any specific framework. With this in mind, I make an effort to provide as framework-neutral as possible conceptions of commonsense and situation holism. These are two human cognitive capacities that are exercised in an open-ended set of circumstances. Under the lights of our cognitive machinery, such circumstances become fully-fledged contexts. A context involves not only the environmental state of affairs, but our expectations, beliefs and goals as well. I them present the chapter's central claim: human contexts are non-saturable. Roughly, this means that we can't predict in advance what kind of feature will play a relevant role in any context. Humans can handle an open-ended set of contexts, and there's no way to draw any clear boundary to the number and kinds of features that characterize it. Consequently, RP presents itself as an explanatory threat in each and every context of human activity. Once these ideas are presented, I use them in order to assess two radically different cognitive frameworks: Fodor's classic cognitivism and Wheeler's "Heideggerian" cognitive science. The point is to show how RP presents itself in both frameworks in a rather similar way, despite one being broadly computational and representational in the most classic sense and the other is non-representational and non-computational. Importantly, this results generalizes to other frameworks, which means you don't get rid of RP by simply picking (say) ecological psychology or avoiding computational mechanisms.

Right before entering chapter 2, a distinction between representational productivity and inferential productivity is introduced. Inferential productivity is about the system's capacity to entertain an open-ended set of inferences. It is the cognitive capacity that requires a solution for RP. In its turn, representational productivity is about what the system has the ability to express through representations. Thus, a central claim of chapter 2 is that representational accounts of cognition face an additional problem, which is that of accounting for representational productivity. Inferential productivity without representational productivity can only accommodate a very marginal explanatory role for representations. But how much of cognition relies on representations is the kind of thing that should be established empirically, not by limitations in one's favorite framework. Fodor's language of thought (LOT) is the classic way in which cognitive science accounts for representational productivity. Unfortunately, not only LOT brings the *frame problem* with itself (here conceived as distinct from RP), it also preempts solutions for it. Structural, non-linguistic, representations are presented as an alternative. Unfortunately, these don't buy us representational productivity, for their expressive power is domain-specific. The proposed solution is to make use of a bunch of domain-specific representational architectures, hence representational *pluralism.* A plurality of architectures enables purchasing representational productivity without raising the frame problem. In particular, I suggest that certain articulations between domain-specific architectures may result in increasing expressive power and that this is a version of what Karmiloff-Smith once called *representational redescription.*

Now, people have suggested using structural representations as a possible solution for the *frame problem* before. But with a few honorable exceptions, most left the suggestion undeveloped. As a result, many authors rejected structural representations on the basis of arguments that can be easily dismantled (e.g. "how can one structurally represent abstract concepts such as democracy?").[2] To avoid the same mistake, and to show a larger portion of structural representation's potential, in chapter 3 I develop their applicability in real world cognitive science. The idea that the mind exploits the representational productivity of structural representations is what I call *representational cognitive pluralism* (RCP). In order to ground the applicability of this idea, I present suggestions for 1) a theory of representational contents that (hopefully) avoids the crucial antirrepresentationalist's concerns; and for 2) the role of representations

[2] Spoiler: you don't have to.

in the constitution of attitude contents like beliefs and desires. Going further, in order to exemplify the applicability and usefulness of RCP's conception of representational redescription, chapter 3 discusses RCP's possible role in the *relational redescription hypothesis* (Penn; Holyoak; Povinelli, 2008a). Very roughly, this hypothesis is as way of accounting for what makes human and non-human animal cognition distinctive, even though they stand on the same evolutionary continuum. RCP's account of representational redescription buys additional dimensions in the relational redescription hypothesis. In such dimensions, we can place more fine-grained empirical hypotheses about what's distinctively human in cognition. The result of this long discussion is a powerful set of representational explanatory tools that do not raise the frame problem and is broadly compatible with the core methodological commitments of most contemporary approaches such as 4EA cognition (embodied, embedded, enactive, extended, affective).

The relational redescription hypothesis is not here just to show how nice (I think) RCP is, though. It helps to draw the larger dimension in which the claims of chapter 4 are going to be made. In that chapter, I make my suggestion of how RP can be handled within RCP. In order to do that, I introduce Cecilia Heyes's *cognitive gadgets* framework and show how it can be used to overcome important limitations in our capacity to learn what's relevant as we learn about the human world. The crucial idea is that cultural evolution can account not just for information that we can grasp with innate learning mechanisms. It can explain how the mechanisms employed in processing cultural information get established as well. This buy us a bootstrap story, i.e. an account of inferential productivity that does not simply assume relevance sensitivity. We can effectively tell the story of how we acquire and store knowledge about what's relevant in the human world. As this story is developed, I show how the presented conceptions of structural representations and representational redescription can play crucial roles in it.

Once we know how we can learn what's relevant, we can concentrate on how this knowledge is stored and used within cognitive tasks. A decisive part of the suggested account is that the world of human activities is not only analyzable into specific situations. It also involves large-scale dynamics that cut across every situation we encounter. Friendship, for instance, comprises duties and roles distributed alongside a temporal dimension. The dynamics of friendship is like a story-template that "tells us" from time to time things like "this is the part where you're supposed to support your friend". Inasmuch as cognition is guided by such templates, it is much like a writer, who has a limited number of plots to choose from, even if the characters and the details of the story are always different. Templates like these, I suggest, are always "trying to happen" within our cognitive machinery. They comprise a sufficiently small yet broad heuristics out of which we can think an open-ended set of unbounded (i.e. non-saturable) situations. The set of available templates, as well as the way we exploit the possibilities they afford is what I call a *cognitive style*.[3] Due to its heavy reliance on cultural evolution,

[3] I know. I'm not a huge fan of the name myself. But it's hard to find a proper name for something that's between the Scylla of an inflexible routine (it may be employed in new ways), and the Charybdis of a completely

such style is a cognitive gadget in Heyes's sense, and hopefully it helps us in understanding how we can avoid RP every time we engage in a cognitive task.

Finally, in the conclusion I consider what has been (potentially) achieved. I also quickly discuss some possible caveats, possible future research and RCP's relation to some contemporary cognitive frameworks.

unconstrained process.

1 THE RELEVANCE OF RELEVANCE

In this chapter I claim that there is a relevance problem (RP) which manifests itself in at least two human capacities essentially connected with human intelligence: commonsense holism and situation holism. To ground this claim, I show that these capacities must rely on the ability to cope with human world contexts, which are non-saturable, raising the question of how we can account for an open-ended relevance-sensitive context productivity. I show how the problem haunts two radically different frameworks: Fodor's cognitivism and the so called Heideggerian cognitive science, which is non-computational and non-representational. From this analysis, I argue that RP manifests itself in all contemporary frameworks. Then I present an overview and set the stage for the positive proposal on how to cope with the problem that I'll advance in the next chapters.

1.1 What intelligence amounts to

What is intelligence? If we try to find an answer by looking at the literature comprising the 1950s cognitive revolution, we'll be disappointed. The only broad consensus we might be able to find there is that an accurate definition of intelligence is a secondary matter. That is, for instance, Turing's position (Turing, 1936). He believed that to define intelligence or thought more generally is a pointless task. All we can do is ostensibly compare someone that does exhibit intelligent behavior (like ourselves) with something else we're not so sure about, like a computer chat bot or maybe certain social network users. This is the basic tenet behind Turing's famous imitation game.

A first consequence was that researchers needed to think about the phenomenon of intelligence through the lenses of their favorite framework or technology. But in choosing a framework, a researcher is inevitably constraining down the set of possible conceptions. Consider two classical examples: 1) intelligence as the capacity to solve formal problems; and 2) intelligence as the capacity to adapt in a flexible and fluid way to distinct environmental contingencies. Both capture something true. We do associate math or logic problem solving with intelligence, and we also associate intelligence with the capacity to adapt one's behavioral outputs to the current circumstances in a flexible, fast and fluid way. The view (1) was prominent among early Artificial Intelligence (AI) researchers and cognitive scientists, which formulated step-by-step algorithms designed to satisfy epistemic constraints. In its turn, the view (2) is prominent among those following the work of Brooks (1991). For him, intelligent behavior is not a product of programs, but something that emerges from the complex interaction of the basic perceptual and sensorimotor capacities of the organism (or robot). Both views have in common that their respective conceptions of intelligence are articulated in terms of the adopted framework, and not the other way around. The shortcomings and difficulties of each approach emerges in trying to achieve what the other does best. On the one hand, logical systems with central processors are incredibly inefficient at handling simple tasks that require keeping track of multiple perceptual cues. To get a product out of the fridge without hitting something in the way is very challenging for them. On the other hand, to claim that one could model the capacity to play chess in terms of interacting sensorimotor capacities is at best a very long shot.

Trying to understand something we don't know in terms of something else we're already familiar with is not necessarily a problem. The strategy is common place in scientific research. What is distinctive about intelligence, however, is that we have no independent description of the phenomenon to be studied. Therefore, there is no neutral target against which framework adepts can measure their progress. In this picture, the only target-related thing that could put a framework down is a problem that's insurmountable yet unavoidable. But how do we know whether a given framework is stuck in one such problem? How much should we try to solve it before start wondering whether it's time to buzz ourselves out of that bottle? Furthermore, even within a given framework there can be divergence on what should be taken as a pretheoretical sign of intelligence. Researchers like Hutter; Legg (2007) have found more than seventy different conceptions, and conjectured that there are as many conceptions as there are researchers. It is true that they all share a significant intersection: the capacity to do the contextually appropriate thing and the capacity to rationally pursuit a goal, for instance. However, even small intra-framework disparities might lead to clashes about whether a given model or theory is acceptable or not, and there is no sign of agreement about what would be a good enough description of the phenomenon that is independent of such models.

Whether intelligence has an essence is still a wide open question, and I have neither the intention nor the tools to answer it. I believe the best we can do so far is to think of intelligence as a cluster of different, yet closely related, cognitive capacities. John Haugeland provides a first clue to outline this cluster. In his view, intelligence is *"(...) the ability to deal reliably with more than the present and manifest"* (Haugeland, 1996, p. 125). However vague, the conception is informative. I'll use it as a starting point and then narrow down its scope towards the specific capacities I want to target.

What does Haugeland mean by "more than the present and manifest"? He has in mind the possibility of being sensitive to patterns or features of the world even in the absence of a direct causal connection. The usual examples come from cognitivist literature, which emphasizes "higher" cognitive capacities like planning a weekend or thinking about what would it be like if someone long gone were still here. Such cases are taken to imply a causal detachment between the organism and the object of thinking. However, at least in humans, this capacity is ubiquitous in "lower" online cognitive processes as well. That's what we do when we take an object to be heavy, for instance. The weight of an object is an unobservable, it is something we deploy to make sense of how that given object causally interacts with its surroundings. Our sense of how fast would the object fall after letting it go in the air is a product of our sensitivity to this non-observable property.[1] This kind of knowledge about weight plays a

[1] What I have in mind are episodes in which one can regard weight as a property that can be attributed to many objects and that comes with certain causal powers. There can be of course less demanding senses in which

role in what underpins both the capacity to predict outcomes ("it will fall if you put it there") and the capacity to infer the likely causes of a state of affairs we happen to find ("I think it broke because it fell from there"). Furthermore, there can be all kinds of intermediate cases among the paradigmatic ones. Typical examples are those in which the causal connection is momentarily lost. If we're visually tracking someone in a crowd from a long distance, we might lose visual for a while, but sometimes we can reliably predict when perceptual contact will be reestablished ("he went behind that group of people, but considering his pace and the direction in which he was going, he'll probably reappear on the other side in a few seconds").

This leaves us with a huge set of cognitive capacities to consider. My intention is to discuss just two that I regard as fundamental ones. Following Haugeland (1998a) I'll refer to them as *commonsense holism* (or simply commonsense) and *situation holism.* Before introducing them, a terminological remark: in order to present my view of these capacities, I'll talk about models. But at least for now I want to be as neutral as possible as to what exactly a model is. Some claim that the organism itself can be regarded as a model of its environment.[2] Others, like the friends of classical cognitivism, might regard them as computational or symbolic. But right now I'm only interested in the capacity to apply a model to the world and use it to ground one's behavioral outputs, that is, in the capacity to render the world intelligible. I take both commonsense holism and situation holism to be characteristic of how humans do it. In what follows, I'll try to describe these phenomena in ways do assume neither properties nor tools from any particular framework.

1.1.1 Commonsense holism

Commonsense holism is about the huge background of knowledge (both know-how and know-that), that underpins human cognitive activity. It's scope might be a matter of dispute, but some examples are probably enough to show that it cannot be regarded as a minor or secondary issue. Consider first some social contexts. If the bell rings at a friend's house, commonsense knowledge is what allows us to know whether it is appropriate or not to open the door. If you're not so intimate, it will probably sound weird or rude. But if you're age-old friends, it might be taken as a favor. These are artificially clear conditions, though. Real circumstances are filled with nuance: if you're visiting your age-old friend whose fiancé has just moved in to his/her house, then despite your previous familiarity with your friend and where he/she lives, it might be more appropriate not to get involved and let the hosts handle it.

Background knowledge about social contexts also underpins other cognitive capacities, such as reading a text. Consider the following story: [3] Rochelle is crossing the street towards a store. As she walks, her wallet falls. Rachel happens to be there and sees everything. She tries

weight is not regarded as a "non-observable". This is will discussed at some length in chapter 3, were we'll address the gap between the cognitive resources available to human and non-human animals.

[2] See, for example, Friston (2013).

[3] Adapted from Marcus; Davis (2019).

to get Rochelle's attention across the street, but to no effect. Rachel then reaches Rochelle's wallet, grabs it and walks towards the same store. Once inside, Rachel approaches Rochelle, they greet each other, and she finally asks: "Aren't you missing anything?". Rochelle frowned for a bit and suddenly widened her eyes. She quickly puts her hand in her back pocket and replies: "Is it with you?". Rachel then gives her wallet back, and Rochelle becomes grateful, for she saved her from a lot of trouble.

This is a very simple story and anyone capable of reading and understanding it is expected to answer questions such as: 1) What kind of trouble has Rachel saved Rochelle from?; 2) Why did Rochelle put her hand in her back pocket?; 3) Did Rochelle knew she had lost her wallet before meeting Rachel?; and finally 4) Did Rachel and Rochelle knew each other? The text has no explicit answers to any of these questions. Indeed, it is striking how little of the situation is explicit. In particular, what allows us to answer "yes" to question (4) is how the story is told. The way Rachel approaches Rochelle is not really expected if they've never met before. We would find it odd to hear that they didn't know one another, and maybe even complain that the text is misleading. If some reader is unable to answer these questions, we're justified in saying that she didn't understand the story. Whenever we write something, we always assume a reader that is capable of articulating what is written with the relevant background knowledge about the social contexts involved. The same knowledge we use when we're actually going through those kinds of situations. Thus, it is important to notice that commonsense holism in reading is not only about figuring out ambiguities or distinguishing serious talking from jokes. It goes much deeper.

What does it mean to say that commonsense is holistic? I use the term to point out that, in principle, all of our background knowledge is potentially relevant to cope with any ongoing affair. If this sounds odd, consider another example, inspired by Samuels (2010): say I'm worried about my dog barking on a Saturday morning, and I'm trying to figure out how likely that is. Then I learn that there's a chess championship going on in another country. Can information about the chess championship be relevant to predict my dog's behavior? However unlikely, if I learn that my neighbor's brother is playing chess at the championship, and that he likes to celebrate with small (but loud) fireworks, that's enough to realize that the chess results might influence the likelihood of hearing dog barks. Thus, by claiming that commonsense is both ubiquitous and holistic, I mean that most of the time, most of our background knowledge about multiple domains of the world is poised to be engaged in most of our cognitive activities.

Holism in this sense is easier to spot in instances of higher cognitive capacities such as those involving linguistic abilities. But that's by no means the only place were this kind of holism can be found. For instance, security staff in airports need to be sensitive to suspicious behavior. But what "suspicious" amounts to? Most of the time, the task requires being sensitive to fuzzy perceptual cues, such as "acting like someone who doesn't want to be recognized". Solving this cognitive problem (i.e. deciding whether someone's exhibiting a behavior that justifies an approach) with above-chance success rate, may require integrating background knowledge from multiple domains. Therefore, this is an example of concrete yet ill-defined problem that claims for commonsense's help in order to avoid an unreasonable amount of errors. It is true that some markers are less fuzzy and can be considered relevant more frequently than others: to appear nervous during a security routine is a classic example. However, even such markers are context sensitive and the security's tolerance for it relies on the current circumstances. For instance, how nervous someone must be before being approached is also a function of how much security staff is available and how crowded the airport is.[4]

Importantly, being holistic is not about actually bringing to the fore as much knowledge as one can at every single task. Rather, most of the time, most of our knowledge must be intelligently ignored. Thus, a crucial element of commonsense is to articulate our knowledge in a way that takes into account the specifics of the situation at hand (or the one being considered). Situations are a way of characterizing what is distinctive about a given circumstance (they're ways of classifying it): it can involve personal aspects (the kind of situation that makes me embarrassed), social practices (restaurants, hospitals, courts...), and even institutional acts (the inauguration of a building that is actually finished). As an example, consider the question "do you know about computers?". An understanding of the situation we're in helps us in delimiting the appropriate knowledge. If we hear the question from a friend that's browsing the catalog of an online computer store, we know that she's after information for buying computers. But if we hear the same question from someone who's having trouble coping with a computer, we're led to conclude she's after information about how to use or fix it. This capacity to modulate the exploitation of one's knowledge according to a situation is a fundamental aspect of commonsense.

Perner's theory of "multiple models" is an account of the human developmental trajectory that can shed a bit more light on what's at stake (Perner, 1993). According to Perner, in its first developmental stage, infants are constantly working with a single world model. In this sense, the infant's world comprises just the here and now. But that limit disappears early on:

> Somewhere around $1\frac{1}{2}$ years (...) infants acquire a series of skills that require facility with multiple mental models. They come to understand means-end relationships, which require multiple models to project the desired state and the necessary steps to get there. They can infer the location of an invisibly displaced objects, which requires extra models for representing past points in the course of displacement. They start to engage in pretend play, which requires an extra model representing the world as different from te way it really is. They learn to interpret representational media, like pictures, language and mirror images, which require models for representing the information conveyed in these media. (Perner, 1993, p. 47)

What Perner calls "model" is pretty close to what we have just referred to as "situation". Indeed, he conveniently dubs the infant who can handle multiple models a "situation theorist".

[4] Importantly, this is not a point about the genesis of commonsense. It might well be that the kind of knowledge integration that characterizes full-fledged human-like commonsense requires language in order to be done. I have no claim about that. The point is just that, once we have commonsense, and however we come to have it, we apply it to both linguistic and non-linguistic tasks. Some ideas of tools that can help in the scientific endeavor to account for the genesis of commonsense will be presented throughout this work.

Thus, Perner is describing a kind of ability characteristic of our commonsense: we can take the world to be in many distinct ways at once, and we apply our cognitive knowledge accordingly.

Commonsense has another property I have not mentioned so far: we cope with things as they happen, and we understand texts as we read them. In Haugeland's words, *"commonsense holism is real-time holism"* (Haugeland, 1998a, p. 50). For every new bit of information we obtain, that bit can, in principle, be articulated with our background knowledge and produce drastic changes in the underlying process. We might be led to apply a substantially different world model on-the-fly, and there's no principled limit for how much of our background knowledge can be deployed in such a real-time fashion. This accounts for the possibility that a word suddenly changes the meaning of an ongoing utterance, or that a new element triggers a significant revision of the whole situation, as when a new clue drives a crime detective to review the role of every other piece of evidence she already has. A solid alibi might suddenly fall. A ruled out person of interest might now emerge as the likely perpetrator, and so on.

At this point, some might be tempted to take commonsense as a capacity to deal with what is typical or ordinary in the human world. In this view, commonsense is just another word for what we're familiar with. Indeed, that's what Dreyfus took commonsense to be in his famous critique of early AI.[5] A satisfactory account of commonsense certainly involves an explanation of such familiarity with what's typical. However, as Dreyfus himself later acknowledged, commonsense goes further. Take this example from his later work (Dreyfus, 1992; Dreyfus; Dreyfus, 1987): a horse racing gambler who's trying to figure out her bet learns that a given jockey had hay-fever and that the race-course landscaping is in full-flower. This is a situation where knowledge about human health must be used to think about what to do in a context of horse betting. Despite involving atypical connections among different sorts of knowledge, we can readily acknowledge its relevance and articulate the role of each piece of knowledge on these grounds.

When Dreyfus presents this case, he readily admits that previous familiarity with what's typical in human wordly activities cannot play the role of a guide towards the appropriate course of action. Nothing in the past of the agent could help him figure out what to do in this situation. Dreyfus is led to conclude that commonsense encompasses a kind of "creativity" which allows humans to render intelligible even situations that are different from whatever it is that they're used to. I'm not sure we should follow Dreyfus in considering this capacity to be a kind of creativity. To be sure, in abnormal enough circumstances, we get lost and have no idea how to cope. But we do have the capacity to handle relatively conservative variations of what's typical, and we do have the tools to articulate pieces of old knowledge in new ways, if the circumstances require so.

In other words, we can extrapolate from what we know by efficient, fast and fluid means. To account for this capacity is a formidable challenge.[6] We seem to extrapolate in a very

[5] See Dreyfus (1972).

[6] Indeed, the attempt to contribute to a better understanding of the way we extrapolate is one of the goals of this thesis.

particular way which no one could completely figure out so far. That's why large language models (LLMs) like OpenAI's GPT family, at least in their current state, can hardly be trusted with tasks requiring human-like commonsense: they can extrapolate from their training *corpus*, but not in a human-like fashion. Very roughly, LLMs work by trying to estimate the probability of a word being used in the context of words that have already been said. Thus, in the context of a sentence like "Hello, how are...", it may guess that the word "you" is the most likely to appear next. But that's certainly not how we choose the words we use. Rather, we seem to rely on reusable world models targeting multiple domains, and we pick our words according to what's rendered by them. For instance, if we have already witnessed a vehicle collision, this knowledge can be applied to estimate what would happen in a collision between trains (or even airplanes), even if we have never seen one. While LLMs may be able to figure out that the vocabulary used to talk about cars can (sometimes) be used to talk about trains or airplanes, relying on estimates of vocabulary-appropriateness is very different from relying on models of causal relations. Therefore, to account for commonsense holism is to account for this capacity as well. Appeals to familiarity with what's typical are not enough.

I hope the examples and the description provided so far are enough to show that having commonsense is a huge and complex achievement for any cognitive system. Insofar as we know, only human animals have it, but I need not and will not commit myself to this claim. Indeed, I'm deeply sympathetic to the idea that other species, mainly those evolutionarily closer to ours, may present the capacity to integrate some knowledge in a commonsense-like fashion. The focus on how commonsense is manifest in humans is a feature of my current target, which is how intelligence is manifest in humans. I think it is an indispensable component of any clustered conception of human intelligence. Thus, commonsense is not to be understood as an all-or-nothing gap between human and non-human animals, but that doesn't mean it is not guilty of some very hard problems in the study of human cognition. Before digging into those issues, I'd like to draw attention to another, closely related yet distinct, cognitive capacity that seems peculiar to the way humans render the world intelligible.

1.1.2 Situation holism

While commonsense holism is already impressive, situation holism raises the bar significantly. We saw how commonsense seems to imply that much of our background knowledge about the world is continuously involved even on apparently simple cognitive tasks. Within commonsense, situations appear as a way of characterizing what is distinctive about a given circumstance (they're ways of classifying it). This leads us to think of the relationship between different situations as one of mutual exclusivity: either you are in situation x, or in situation y. But humans (and perhaps other creatures) can do much more than relying on different types of situations when taking the world to be in a certain way. We can also articulate them, coherently working out indefinitely many distinct patterns of interdependence. That's what Haugeland dubbed *situation holism*. In order to see what's at stake, consider what it takes to read and understand the following story:

> One evening, the Khoja looked down into a well, and was startled to find the moon shining up at him. It won't help anyone down there, he thought, and quickly he fetched a hook on a rope. But when he threw it in, the hook snagged on a hidden rock. The Khoja pulled and pulled and pulled. Then suddenly it broke loose, and he went right on his back with a thump. From where he lay, however, he could see the moon, finally back in the sky where it belonged — and he was proud of the good job he has done. (Haugeland, 1998a, p. 53)

In order to grasp Khoja's pride and sense of achievement, one must keep track of how two distinct viewpoints interact: the "real" one, and the epistemic one, i.e. the one in Khoja's imagination. Each one accommodates a different situation.

But what's at stake in situation holism is not only the specifics of each situation taken in isolation. It is rather about how the development of a situation being tracked might prime or constrain the others. In order to properly understand the text, we need to keep track of how these situations interact. In other words, we need to grasp the underlying plot and understand why the way these distinct viewpoints interact matters to the story being told.

> This demand for over-all coherence — that all the various "situations" (with respect to which clauses are understood) should fit together in an intelligible way — is what I call situation holism. It is a general feature of natural-language text, and coping with it is a prerequisite to reading. (Haugeland, 1998a, p. 54)

What's remarkable concerning this example is how little about that interaction is explicit in the text. There is no clear reference to the Khoja's misperceptions, and yet most of us readily see that "the moon shining up at him" is just the reflection of the moon in the well. If a reader cannot get that, we'd say she did not understand the text at all.

In the resulting picture, commonsense is involved: we known that the moon won't fit within the well, for instance. But (pardon the pun) commonsense can't tell the whole story. It plays a role in recognizing the distinct viewpoints (or distinct models, in Perner's words) as alternative takes on the Khoja's world, but situation holism emphasizes their articulation into a somewhat unified, integrated whole. Situation holism accounts for the continuous trade-off between them, i.e. the interactions that take place as the narrative develops.

Situation holism has been studied in communicative contexts by relevance theorists such as Sperber; Wilson (1995). However, it is not an exclusively communicative phenomenon. While it is true that Sperber and Wilson also make claims about cognition in general, and not just about language processing, I don't think their approach is enough to account for situation holism when conceived as a problem about cognition.[7] And how should we understand situation holism at the cognitive level? Just like commonsense, situation holism is poised to play its role whenever we handle the structure of human activities in the world.

[7] Their approach will be properly discussed shortly.

As another example, consider children engaging in pretend play. They can gracefully switch between two or more pretend scenarios (one in which the piece of wood is a chocolate, and the other in which it is a cell phone), and they do that without loosing track of whatever function the piece of wood has in non-pretend scenarios. The complex interaction between the way an object is regarded in a pretend scenario and its identity in the real world becomes manifest when the object gets broken in virtue of an event that imposes itself from outside: the piece of wood might get wet, the chair regarded as the king's throne might make some funny noise or even break. In circumstances like these, pretend participants must choose whether (and how) they'll accommodate the events in the play. The chocolate bar got wet and started to melt, so the children now proceeds to pretend-wash their hands. The funny noise may be just ignored, but perhaps the king becomes angry and requires another throne. Something similar can happen if the chair gets broken: the king might require a new, better one, or they can just preempt the play and start to worry about what their parents are going to think when they find out. In a nutshell, they must decide when to enforce a rule of the game and when to resume real-world inferences. This is the kind of decision-making that relies on situation holism.

These examples also show that, just like commonsense, situation holism cannot be regarded as a mere familiarity with what's typical nor with a collection of previous experiences. Whatever underpins the children's behavioral outputs in pretend play (as well as our own when dealing with the huge network of institutions and structures comprising the human world), it can also be effectively applied to novel situations. Situation holism also seems indispensable to the cluster of capacities that comprises human intelligence.

We have now a framework-neutral characterization of both commonsense and situation holism. Before introducing the kind of challenge that emerges when trying to account for them, there's one last crucial idea to introduce: the non-saturable character of human-world contexts.

1.2 The non-saturable character of human contexts

Most of the difficulty in explaining commonsense and situation holism is connected to their reliance on context-sensitivity. Commonsense relies on the capacity to attribute the circumstantially appropriate weight for each portion of our background knowledge about the world. Likewise, selecting the circumstantially appropriate way to articulate two (or more) models of a given situation is necessary for situation holism. But why and how exactly is context-sensitivity an issue for cognitive science?

In order to answer that, we need to rely on a good description of the phenomenon. I have no intention to provide a complete assessment of context, for this would take us too far afield. However, I think context, inasmuch as it characterizes the background in which human behavior is embedded, has properties that are not shared by all cognitive creatures (perhaps none). In what follows, I'll present what I mean by "context" in general and "human context" in particular. But before we do that, some preliminary remarks are in place. First, as I use the term, it is restricted to creatures with some kind of cognitive machinery. Consequently, different takes on what cognition is results in distinct sets of creatures. I think many species of animals have what we should regard as cognitive capacities, but any disagreement about this can be easily put aside, for the core target of the discussion will be context in human agents. Second, different cognitive frameworks afford distinct characterizations of context. The specifics of some frameworks will be extensively discussed throughout the chapter, but right now, the target is a somewhat crude and framework-neutral conception. Third, I don't want to commit to any underlying conception about the make-up of the world. Whether the world comprises facts, objects, or what have you, bears no weight on the discussion. Finally, it must be crystal clear that by characterizing human contexts and contrasting them with those of other animals, my intention is not to describe an unbridgeable gap. Rather, I want to distinguish two opposite poles of a continuum that makes room for intermediate cases. Thus, it is not only possible, but rather likely, that we can find proto-versions of human-like contexts in animals that are closer to us in evolutionary time. How close they can be (when it comes to context) is still a wide open question, though.

How do we go from states of affairs in the world to contexts? Consider first a simple non-human animal like a tick. Just like any physical entity, the tick is subject to the causal effects of other entities or forces. The precise effect that any causal impingement has depends on the nature of the impingement and on the make-up of the creature's body. The sun's warmth may have a bear on its body in a way that the sun's light has not. A falling stone might hurt it in a way that affects its normal biological functions, or might be an event with no relevant effect in the animal's body. Within the set of possible causal influences, there is the subset of those that can bear on the creature's behavior. Such influences are also stimuli. Even the most rudimentar stimulus can be regarded as a kind of perceptual cue the animal can exploit to produce adaptive behavioral outputs. The set of features to which the creature is sensitive in this way constitutes its *effective environment*, which is roughly what Von Uexkull (1934) called *Umwelt*.[8] Different creatures inhabit different effective environments, and they cope with the current states of affairs only through their lenses. The tick, for instance, is specially sensitive to the butyric acid that can be found in the skin of mammalians. The acid is among the few things that can elicit some kind of behavior from the tick or, we could also say, it's among the few things that "exist" for it. Thus, the effective environment is the world's state of affairs as perceived by the creature.

Effective environments yield two dimensions on which robust behavioral outcomes can get stabilized. First, it provides cues an organism can use to determine the right time to deploy some behavioral strategy. By following this road, even a very simple "cover-all" strategy might

[8] The name "effective environment" is from Clark (1997a). Those already familiar with Robert Cummins' conception of cognitive targets (Cummins, 1996) can help themselves by understanding the creature's effective environment as the set of possible targets it is able to fix on. Readers not familiar with this idea need not worry, though. For our current purposes, it is nothing but a shortcut. We won't really need it until chapter 3, where it will be properly introduced.

become adaptive. Flowers, for instance, simply spread their seed in every direction.[9] This works reasonably well because they're sensitive to seasonal cues indicating the right moment to carry out the task. Ticks use a similar strategy. They can hang idle on the tip of a tree branch until some cue indicates a passing mammalian, giving it the chance to let go of the branch on the right time and drop safely on mammalian skin. But there's a downside for invariant behavior: the adaptiveness of such strategies relies heavily on the effective environment's stability. A radical change of circumstances can suppress relevant cues, perhaps leaving the creature in an endless idle state. It might also render the cues useless, leading to harmful false positives and causing the tick to trade its branch for non-mammalian skin, for instance.

As for the second dimension, effective environments also allows creatures to achieve some adaptive outcome robustly by behaving differently as a function of the available cues. Some predators can rely on lots of different cues, both from their prey (scent, visual characteristics, and so on...) and their surroundings, in order to select not just the right moment to hit, but also the best way to approach their prey. Needless to say, nature treads on both roads. Animals usually employ a mix of different strategies. Consider the impressive navigational capacities of some species. Salmon are sensitive to the scent of their natal stream. When the time to spawn comes, they can use this scent to identify their home stream even after spending years in the ocean, which means there is a rudimentar kind of mnemonic mechanism involved. But that doesn't exhaust its tool set. Scientists also believe that it is sensitive to the earth's magnetic field, which accounts for its capacity to search for the river they came from. Once in their home river, they start relying on scent to identify their home stream. The application of (at least) two distinct mechanisms is a step towards alleviating the reliance on the stability of its environment. This can be generalized: while "cover-all" strategies can be powerful, a well-endowed set of behavioral strategies can go further.

Consider now a more complex creature with some kind of cognitive machinery, like an ape or an elephant.[10] They have a richer effective environment, at least when compared to ticks and salmon. But that's not the only difference. Their cognitive apparatus allows them to have a take on their effective environment by rendering it intelligible somehow. This rendering works like an intermediate framework between the creature and its effective environment. Cognition involves internal states and processes that bear on the creature's behavioral outputs.[11] In a very liberal sense, we can say that those states and processes characterize the creature's expectations while coping with its effective environment. In light of these expectations, the creature's view of the environment changes. Some perceptual cue that would usually trigger a certain behavior gets inhibited, for instance. Furthermore, new possibilities emerge out of

[9] This example is from Shea (2018).

[10] I take the presence of cognition in apes and elephants to be noncontroversial, and I'll ignore behavioristic complains at my own risk.

[11] Importantly, by "intermediate" I don't mean to postulate some entity inside the cognitive machinery that stands in for what's outside, i.e. I'm not making a representational claim. Representations are just one among the available tools we can use in trying to account for such intermediates. One could also, for instance, use gibsonian language and take them to be the set of abilities employed in handling affordances (Gibson, 1979).

the complex interaction between the effective environment and the creature's expectations. This means that the creature can be sensitive to new features, for instance, non-perceptual features, articulations of previous experiences ("the stick I've just saw can be useful to reach this fruit") or its own predictions about what's going to happen next, given the current state of affairs (e.g. a rat in a maze may predict the presence of food in a certain place at a certain time). Consequently, its effective environment becomes larger, enabling increasingly complex behavior. That's what I take *context* to be: whatever the effective environment becomes under the creature's cognitive lights.[12]

Context is thus relative to the agent's cognitive capacities. This means that, as creatures endowed with cognition, elephants and apes exercise their capacities inside contexts, not just effective environments. But what about ticks? Given the likely absence of cognitive activity, should we say that its context just happens to be identical to its effective environment? We definitely can, and those advocating for a conception of cognition that's broad enough to encompass whatever accounts for the tick's behavior definitely must. Here, however, I want to make room for the assumption that there is some distinction between cognitively and non-cognitively produced behavior, whatever that may boil down to. Particularly, I think that cognition is characterized by the capacity to transform sensory inputs in knowledge about the effective environment, allowing the creature to make use of that knowledge in producing its behavioral outputs.[13] Notice that this conception is fully compatible with the idea that our effective environments play a constitutive role in our cognitive capacities. As I see it, this would be just another way of claiming that cognition is embodied and embedded. But it can also accommodate a more classic and "detached" view of cognition, such as those who see the body and the environment as nothing but a source of inputs and a place to throw outputs. Though I personally think that cognition is embodied and embedded, the current point is not to discuss which conception better accounts for the nature of our cognitive capacities. What matters is to notice that, as far as the distinction cognitive/non-cognitive makes sense, so does the distinction between context and effective environment, even though one could eventually disagree on the adopted terminology. Therefore, context becomes closely related, yet distinct from effective environment. That's why two individuals may share an effective environment without sharing a context (they might have distinct understandings, goals or expectations, for instance), but they cannot share a context without sharing an effective environment. The size of the gap between context and effective environment depends on the creature's cognitive capacities. For instance, apes probably cannot make use of abstractions when coping with their environments, at least not like humans can.[14] Thus, even though humans and apes can share

[12] This does not mean that context it is subjective as in private. The creature need be neither conscious nor potentially conscious of whatever is going on in its own cognitive apparatus. Moreover, there's no fundamental issue precluding the possibility of fully describing a creature's context from a third-person perspective.

[13] I'm using the word "knowledge" in the same liberal sense we usually find in cognitive science's literature. It is compatible with both know-how and know-that. It also goes without saying that this conception of cognitive knowledge is distinct from the one studied by epistemologists. What's at stake is not the status of some full-fledged agent's attitude, but whatever gets stored in its cognitive mechanisms.

[14] This claim relies on the work of Povinelli (2012). It will be discussed at some length in chapter 3.

effective environments, the contexts in which our actions take place would still be distinct, given the discrepancies in our cognitive capacities.

As another example of this point, consider dogs. We share a significant amount of the same effective environment. That's what allows us to regard them as pets. There's a considerable intersection between what we can hear, feel and see, to the point that we can interact by, say, throwing a tennis-ball and expect it to be tracked, chased and brought back by the dog. However, the fact that we don't share the same cognitive machinery means that we won't handle our effective environments in the same way, which means we'll perform within distinct contexts. For instance, they might see a tennis-ball as a tennis-ball, but only in the sense that it is a thing with distinct physical properties like shape, color, texture, and so on. On the other hand, we can also see tennis-balls in a functional, game-embedded sense. This is why we could regard even balls with distinct physical properties as tennis-balls, provided that they satisfy some game-enabling standards, such as size, elasticity or weight. Thus, while we might never be able to share much of a context with a dog, we can nonetheless share a good deal of our effective environments. In contrast, the lack of such an intersection makes it really hard to have bats as pets.[15] Our effective environments are just too different. Bats can track lots of things we can't and vice versa. We have nothing like an echolocation system, and even when it comes to the kind of senses that we happen to share, like vision, our sensitivity to the environment is mostly disjunct, for bats can see in conditions we would consider pitch black.

Furthermore, effective environment and the available cognitive machinery can play distinct roles in our endeavor to learn whether a creature has a given capacity. The distinctiveness might not be easy to see at first because there's always the risk of conflating them when trying to guess which cognitive capacities can be attributed to a species. I believe this is what Frans de Waal has in mind when he asks whether we are smart enough to know how smart other animals are (De Waal, 2016). Elephants are a nice example: for years, scientists thought that elephants could not use tools. They are known for using their trunks to grasp things, and so researchers formulated experiments in which the animal was expected to grasp a stick with his trunk and use it to reach a reward. As elephants consistently refused to do so, many took it as a sign that the animal didn't get the problem, which is a claim about the creature's cognitive capacities. However, Frans de Waal describes how Kandula, a male elephant living in a Washington zoo proved this assumption wrong. In trying to reach a tasty reward high above his head, Kandula quickly noticed that he could kick a wood box with his foot until the box was underneath his reward. He then stood on the box with his front legs and finally managed to reach the reward with his trunk without the use of anything like a stick. The issue is that the elephant's trunk is also its nose. In grabbing a stick, the nose is blocked, which preempts the animal from tracking its reward. What was previously regarded as an incapacity to use tools proved to be our difficulty in recognizing the role of the animal's distinctive effective environment. Elephants can use tools in surprisingly smart ways: the box used by Kandula

[15] Thanks to Ernesto Perini-Santos for the example.

was purposely left far from the reward, so he had to take distance and seek for the box before coming back with it. This kind of example exposes the risk of consistently misunderstanding the animal's effective environment and thus failing to evaluate its cognitive capacities.

Given the stipulated definition of context, it follows that whatever is distinctive about human contexts must be grounded in something distinctive about human cognition. Somehow, humans can go far beyond their effective environment and extend it in a largely unconstrained way. But where exactly lies the difference? There's obviously no framework-neutral answer to this question, for any claim about the nature of the underlying processes will be commited to some cognitive framework. There is, however, a fundamental (and framework-neutral) difference in the way we can conceive human contexts. It regards the possibility of saturation. Consider another example related to animals' navigational abilities: the birds' capacity to migrate. It involves considerably complex mechanisms. While studying white-crowded sparrows, Thorup *et al.* (2007) have shown that the behavior of juvenile first-time travelers and adults is different because the latter is sensitive to features unavailable to the former. The adult's previous travel experience makes a difference for the chosen direction. At some point he became sensitive to new cues, perhaps by learning. This suggests that the bird's behavior might rely not just on different features, but also on complex articulations of them. We can thus picture the context in which the bird's behavior takes place as a certain articulation of a subset of the features to which it is potentially sensitive to, given its cognitive capacities.

Despite its potential complexity, the bird's context is always *saturated*, i.e. there's a fixed number of features it is sensitive to while in that context, and a fixed role for each of them to play as well, which means that there can be an ultimate description of the feature's subset in question.[16] An ultimate description of a context is one in which 1) adding new features (or new ways to articulate them) will make no difference to the creature's behavioral output; and 2) removing a feature (or a possible articulation of them) will necessarily make a difference. To unpack what this means, let's put the bird aside for a moment and consider a toy-like example of a cognitive system that is sensitive to four features: a, b, c and d. It is not sensitive to all of them all the time, for it depends on its current context. Say the system can be in up to three different contexts.

Table 1 – Toy-like cognitive system.

c1	c2	c3
a	-	a
b	-	-
c	c	c
-	d	d

[16] The point is not epistemic, i.e. what's at stake is not how much of the bird's context can be known to us.

Thus, by reading the table we can see that when the system is in context c1, it is sensitive to three features (a, b and c), when in c2, it is only sensitive to two of them (c and d), and so on. Importantly, what the current context stipulates are the subset of features to which the system is sensitive to. It does not mean that the features are necessarily present in the system's environment. What it does mean is that, while in (say) c2, the eventual presence of feature d will have an effect on the system's behavior. In the same vein, the presence of feature d while in c1 will bear no effect on the system (it will simply keep doing its business as if d wasn't there). It may trigger some action or some change of perspective (such as switching to c1).

What is lacking from the table is the way in which a context articulates the features it is sensitive to, and every context necessarily involves some articulation. In this sense, being in c3 is not just about being sensitive to features a, c and d. It is also about what to do with them. Thus, c3 may comprise an articulation in which the system is led to stay on the lookout for any of them simultaneously, while c1 may comprise an articulation in which it may lead the system to actively look for a, start looking for c only if it takes too long, and though it won't look for d, it will immediately stop searching for a and c in case d appears. Crucially, the way the system switches from context to context is also a given of how every context articulates the features it is sensitive to. Thus, if the system is in context c1 and the presence of a, b and c lea it to c2, this is part of the specification of c1. For instance, if there's no determined path from c1 to c3, that means the creature can never go from c1 to c3. It also means that, if the only way for the system to leave c1 is through the presence of a, b and c, it will be stuck in c1 as long as it takes. In a nutshell, if we provide the way that a, b and c are articulated, as well as the behavioral effects they have in the creature while in c1, we'll be formulating what I've dubbed the ultimate description of c1. Evidently, the same goes for c2 and c3.

At this point, it is important to notice that what underlies the claim that the toy-like system's contexts are saturated is not their low complexity. Chess engines are pretty complex systems, with a huge number of possible contexts in which lots of features are articulated, and yet all of them invariably work with saturated contexts as well. Rather, the heart of the matter is that the system is unable to extrapolate from its set of possible contexts: it may be huge but it's not open-ended. Thus, even if we endow it with the capacity to learn how to be sensitive to new features (or to render new articulations of them), the saturated character of its contexts remains intact.

With this in mind we can go back to the bird's example. Though the bird's ability to migrate involves learning, it does not preempt its contexts to be characterized as saturated, for the learning is presumably specialized (it requires specific experiences and renders sensitivity to specific features). When the bird engages in migratory behavior, it's expectations are such that some features will have no effect. In other words, it is in a context in which it is sensitive only to a certain subset of the features it can handle. For instance, if the sunlight is a feature that makes no difference for the bird's while in its migratory context, then it is simply not part of the context in which the action takes place. That is the case even if the sunlight makes a difference in (i.e. can be a part of) other contexts. This stipulation implies that if a given feature is absent, nothing can take its place (for nothing else makes any difference). That is, while in the context of migratory behavior, nothing can take the place of the feature it uses to guide the direction in which it flies. That's what happens if the relevant sources of information for migratory birds are absent, as in the young white-crowded sparrows' case. Furthermore, a very similar reasoning can account for the cognitive flexibility we find in the example of the elephants capacities.

At this point, a potential confusion must be precluded: wouldn't the addition of some non previously specified feature to a saturated context amount to a change of context? No. Remember that context is the current state of affairs from the creature's perspective. From our God-like theoretical vantage point, we surely can hypothesize about the effects that the addition or substitution of a feature could have, but that's a different matter. If a feature has the potential to dislodge the current context to another one, then that role is already part of the context's full specification. Context-switching is just one among the possible effects that a feature can have in a given context. As an example, consider again the tick. It is not a cognitive creature, but the simplicity of its behavior will be useful in making this point. Assume it's idling in a branch just waiting for a beacon of butyric acid. Even if the tick is potentially sensitive to cues other than butyric acid under different circumstances (such as when it is already enjoying mammalian skin), while in "waiting-for-mammalian" mode, none of those cues will have any effect on its current mode of operation.[17] Likewise, in cognitive non-human animals, the pathways between the current context and all the other possible ones (given the creature's cognitive capacities and effective environment) are given by the current context. In more impressionistic terms, the full description of saturated contexts enables a kind of leibnizian *calculemus* of the animal's behavior, for there's a clear finite basis for an equally finite set of behavioral outputs. In the end, this is what allows us to conceive non-human contexts as subsets of features.

In contrast, human contexts' are *non-saturable*. Non-human animals can perform actions in complex environments, for sure, but nothing close to the productive and flexible way in which human contexts can be characterized (from a theoretical perspective). While handling our effective environments, creatures like us need not just cope with a large set of possible circumstances, but with an open-ended set of them. Consequently, there's no ultimate description of human contexts, for it's always possible that adding or substituting a feature (or a way to articulate them) makes a difference. In order to start unpacking what this means, consider our navigational capacities: even if completely deprived of our usual navigation tools (such as knowledge about the surroundings) we can use the earth's magnetic field through a compass, the movement of the sun and even the seasonal course of constellations. It's true that not many of us actually know how to use any of those features in order to navigate (I certainly don't), but the point is that we could learn to become sensitive to these and other related features (even if only with the help of some gadget). In this sense, maps, gadgets and

[17] Perhaps I should say philosophical-tick, for the description we're working with here is an obvious oversimplification of the creature's actual mechanisms.

written instructions have the potential to become salient navigational cues.

But a similar reasoning applies to every human context. Let us consider a relatively common one: Alan and Bob are friends at a bar, and Alan wants to have a beer. The tough cognitive task that arises is deciding whether he should invite Bob to join him. If Alan were a cognitive system capable of dealing only with saturated contexts, each situation would need to be described through a subset of features, just like the previously described toy system. Let us start with what appears to be the simplest one:

(3) Alan and Bob are at a bar late in the afternoon.

In this context, should Alan offer Bob a beer? If these are the only features being considered, there seems to be no reason not to invite him. But consider now what happens if we add another feature.

(4) Alan and Bob are at a bar late in the afternoon. They're going back to work right after.

Suddenly there seems to be a plausible reason to hold back the urge to have a beer. Perhaps an additional feature can be of help:

(5) Alan and Bob are at a bar late in the afternoon. They're going back to work right after, but the work they'll engage in is not really that complicated.

If the work in which Alan and Bob will have to do late at night is not really that demanding, then perhaps a beer won't hurt, which means Alan may have its beer and it is plausible to invite Bob to join him. But consider now another feature.

(6) Alan and Bob are at a bar late in the afternoon. They're going back to work right after, but the work they'll engage in is not really that complicated. Bob has a long-standing alcoholism problem and has been sober for two months.

Is it OK to offer a beer to someone with a history of alcoholism who is trying to stay sober? It seems like Alan will have to leave it for another day (and perhaps with another company). But lets add yet another feature.

(7) Alan and Bob are at a bar late in the afternoon. They're going back to work right after, but the work they'll engage in is not really that complicated. Bob has a long-standing alcoholism problem and has been sober for two months. On a previous occasion Bob has said that he doesn't mind when people drink in front of him.

It seems like Alan is going to have his beer, after all. The point of these examples is (hopefully) clear enough: there is no stopping adding new features. There is no (non-arbitrary) point in which the context is "characterized enough" in the sense that there's no other possible feature that can make a difference in Alan's behavior. We can imagine all kinds of possibly relevant additional characterizations: for instance, the fact that the temperature might be too high and there's no air conditioning might make one change his mind about what to drink. Thus, no matter how rich is the description of a context, the description is never able to exhaust (i.e. to saturate) all the possible features one can attend to in order to decide whether it is OK to have a beer with a friend. No matter how complex the current context is, there's always the possibility of realizing the potential role of another feature.

To make it even worst, this is not only about the possibility of indefinitely multiplying the number of features within a context. There are also indefinitely many ways to articulate the already considered set of features. To see this, consider another tough cognitive task: some friends go to a restaurant to eat a *moqueca*, but once there they find that it is unavailable at the moment. Should they pick another meal ou should they go to another restaurant? Consider these descriptions of the situation:

(8) Some friends are going to a restaurant in order to celebrate the birthday of one of them. They intend to present him with his favorite meal: *moqueca*.

(9) Some friends are going to a restaurant in order to celebrate the birthday of one of them. They intend to present him by paying the bill.

If we understand the situation as described in (8), then it seems like changing the restaurant is appropriate. But if we understand it as described in (9), it might be considered a bit of an overreaction. And again, a similar reasoning applies to all and every human context.

Furthermore, human contexts can interact with each other in complex ways. There are dinners in hospitals, medical emergencies in restaurants and both at the same time in airplanes. Thus, when characterizing a context, features and expectations regarding restaurants may interact with features regarding friendship, work goals or hospitals. How can we, as theorists, map all and every possible pathway among the set of possible contexts of a cognitive system? The bitter answer is that we can't.[18] Consider the shifting from a context where one is having dinner and then a colleague shows up. Where should we go from there? Is it good to have someone from work around, or is it bad? If it's bad, how bad? Is it a minor distraction, does it produce some level of anxiety or is it a trigger to get out immediately? Again, we face a set of possibilities that's prohibitively large. The point is that, in describing a non-saturable context, we can't model the absence of a feature as the absence of an element in a subset. The absent feature can play a significant role not just in choosing a beverage, but also in context-switching, i.e. in characterizing the context that results from whatever changes. In other words, even absent elements (like work colleagues) might have a role in explaining what we end up doing (or drinking). In this picture, the conception of context as a finite subset of features must be put aside. Human contexts are better conceived as a holistic rearticulation of most (if not all) of our cognitive world, i.e. a context is a more or less global state in which the overall set of features we're sensitive to are more or less salient in different degrees. In more palatable

[18] The grounds for this claim will be further developed in the upcoming discussion.

words, a context is a species of "tonality" that the world acquires. Some features shine, some become gray, but none completely fades to black. From now on, this is what I'll refer to as *contextual tonality.*

What kind of cognitive capacity can underpin the ability to navigate non-saturable contexts? Non-saturability cannot be understood as the ability to process context descriptions with an unlimited number of parameters. That would only make sense for creatures with limitless time and memory. Rather, non-saturability captures the fact that there's no way to know in advance what are the kinds of features or parameters that we'll have to cope with in any given situation. What underlies it is a kind of productivity. Such productivity is not like compositionality, usually associated with linguistic capacities. Instead of the capacity to render increasingly complex compositions of primitive elements, it comprises the capacity to articulate the available cognitive resources in indefinitely many ways and generalize or extrapolate from these articulations. As this capacity can be recursively applied to whatever context we find ourselves, there is no context that we cannot extrapolate from, i.e. there is no context that we cannot rearticulate with some feature that was being ignored up to that point.

Importantly, "rearticulating" doesn't necessarily mean "adding a new feature as salient" as if we were adding a new dish to a stack of dishes. Rather it might also mean changing the role of one or more features, which is like to change the arrangement of the dishes, or maybe the order in which they are stacked. That being so, what's striking about this capacity is not that it implies non-natural capacities, for memory and time constraints do apply. If we rearticulate a context c1 into c2, and further rearticulate it into c3, c4 and so on, it certainly there is a point at which we lose track of the trajectory, and so maybe at context c35 (or c50) we may leave out some feature that was salient in c1. Therefore, what requires explanation is how we can always extrapolate from any contextual tonality we find ourselves in.

Indeed, as we'll see in the forthcoming discussion, accounting for the non-saturable character of human contexts is a major challenge for the cognitive sciences. But before digging into that, we'll explore a bit how the explanatory role of context was conceived in some well-known cognitive frameworks. This will help in at least two ways: first, by throwing some light on the many pitfalls ahead. Second, by making salient how easy it is for us to let slide some essential features of context as a cognitive phenomenon.

1.3 The role of context in cognitive explanations

There is no doubt that context-sensitivity is an achievement of the mind. In this sense, contexts are among cognitive science's *explananda.* As theorists, we want to explain the capacity of a system to behave properly within what we characterize as a context. But could some notion of context also play a role in explaining our cognitive capacities? The idea that capacities manifest in full-fledged agents can inspire a theoretical posit for their internal machinery is not new. The cognitive revolution in the 1950s comprises one such move: in the cognitivist picture, the use of representations is an achievement of the mind (we use maps, graphs and language as stand-ins all the time) that inspired a theoretical posit that allows us to explain how the mind's mechanisms work. Could context *qua* cognitive capacity of cognitive systems be a similar inspiration for some kind of theoretical tool (a domain, a process, a structure, what have you) that has an explanatory role to play in scientific accounts of the systems' underlying cognitive mechanisms? Notice that the question is not whether the system has personal-level access to what characterizes the context. Rather, what's at stake is the possibility of a move from a notion of context *qua explananda* to a notion of context *qua explanans*, both from a third person perspective. We have just discussed what it is for a full-fledged agent or system to perform an action inside a context. And we can understand what it means to say that the agent's actions take place in a (say) restaurant context: it means that the agent's behavior respects certain behavioral constraints that the context implies. But could some still-to-be-unpacked notion of context be useful to explain this very capacity? A notion of context that could participate in explanations of the mind's mechanisms? In order to avoid confusion, let us call this potentially explanatory notion *m-context*. What could it possibly buy us explanation-wise that we can't get without it?

Let us start by considering an example of what a system that does not rely on any notion of m-context could work. Say you're idling alone at home during a COVID-19 quarantine. Suddenly the door-bell rings and you're now trying to decide what to do about it. This brings about a change of context. Up to a few seconds ago you were in a "home alone in quarantine" context, and now your in a "home alone in quarantine with someone at the door" context. How should we account for this change of context within the mind's mechanisms? In approaches inspired by folk-psychology (as well as classic AI), the world's current state of affairs is designed as a pool of belief-like sentencial entries that describes it (or more precisely, a pool of language-like representations of such belief-like contents).

Say you're one such system. Whatever is in the pool is taken by your mind mechanisms as something true about the world, and thus they operate with that information in order to produce your 's behavioral outputs.[19] In this case, the change of context could be accounted for by simply adding a new entry to the pool of belief-like entries. Something like "doorbell ringing now". This changes the set of possible inferences that can be made out of the pool, and thus may result in new behavioral outputs. Thus, the simplest possibility is to claim that you've just acquired a new entry in this cognitive knowledge database of yours, and that whatever you make of it is grounded on the inferences you can make from this new entry. In this account, there is no cognitive effect beyond what comes from the acquisition of the new entry. But if we can account for the resulting behavior simply by pointing out the presence of an entry, what's left to be explained? At first, there seems to be no role for any conception of m-context to play in the explanation of the resulting behavior. Thus, in order to claim that there is a non-trivial role for m-context in mechanistic accounts, one has to 1) show that the mere presence of a knowledge entry in the system is not enough to account for the agent's

[19] Remember that personal level access to that information is not what's at stake and comprises another matter.

behavior in the new context; and 2) show how some conception of m-context can account for whatever is lacking.

An answer for both questions that we can frequently find in real-world frameworks of cognitive science is this: m-context plays an important role in the overall organization of the system. This is not the time to discuss whether they're good or bad approaches, though. Our current target is just to identify what they take m-context to be, and how they can have a nontrivial role in explanations of cognitive capacities. The first example comes from AI's knowledge representation literature: an m-context is a data structure.[20] The main concern found there is how to organize information in a way that allows for efficient retrieval. Failing to do so would result in systems that could get lost in a messy ocean of knowledge. The m-context's role is thus to organize the agent's inner cognitive life and avoid this. This is why simply adding another entry to an unstructured pool of belief-like entries is not enough to account for the system's behavior. Minsky's frames (Minsky, 1997) and Schank's scripts (Schank; Abelson, 1977) are probably the most well-known approaches. Both relied on heavy data compartmentalization. In the now classical example used by Schank, there could be stereotypical scripts for restaurants, hospitals and so on. Inside those scripts, one could find all the relevant knowledge about how to behave in the corresponding human-world situation. In this picture, the structural resemblance between context and m-context (i.e. the way information is compartmentalized) partially explains the system's performance: the system can behave adequately within a restaurant context because the involved mechanisms are influenced by a restaurant m-context, i.e. by a script. A different organization could seriously affect the efficiency of the system, for it could get lost in processing information unrelated to the current context. In this picture, it is clear that the mere presence of some new piece of knowledge is not enough to explain the system's behavior, for the way these are stored and organized also matters. Essentially, this approach tries to vindicate the commonsense agential level understanding of context by taking m-context to be a data structure that models it. A kind of folk-context, if you like.

However, the notion of m-context is not restricted to classic, computational approaches. An m-context can carry the same explanatory role even in an embodied, embedded and not necessarily representational approach. As a second example, take the framework advanced by Wheeler (2005), which will be discussed at some length later.[21] For Wheeler, the mind is essentially a collection of *special-purpose adaptive couplings* (SPAC). By using the word "coupling", Wheeler emphasizes the fact that the mechanism is embedded. In order to see how this makes room for m-context, we can take Wheeler's favorite example: the cricket phonotaxis system as described by Webb (1993). The mechanism can be found in female crickets and its function is to find male mates through the tracking of the very specific auditive stimulus they

[20] The use of an AI framework is justified by the fact that in the 1970s and 1980s cognitive science and AI were still closely coupled enterprises.

[21] I say "not necessarily representational" because, despite the fact that Wheeler's approach does reserve some space for representations in cognitive explanations, his conception of a sub-agential role for m-context does not rely on it.

produce Here is how Wheeler describes it:

> The basic anatomical structure of the female cricket's peripheral auditory system is such that the amplitude of her ear-drum vibration will be higher on the side closer to a sound-source. Thus, if some received auditory signal is indeed from a conspecific male, all the female needs to do to reach him (all things being equal) is to continue to move in the direction indicated by the ear-drum with the higher amplitude response. So how is it that the female tracks only the correct stimulus? The answer lies in the activation profiles of two interneurons (one connected to each of the female cricket's ears) that mediate between ear-drum response and motor behaviour. The decay rates of these interneurons are tightly coupled with the specific temporal pattern of the male's song, so that signals with the wrong temporal pattern will simply fail to produce the right motor-effects. (Wheeler, 2008, p. 334)

According to Wheeler, SPACs work correctly only in the presence of the right stimulus. There is no need to parse or "interpret" the auditory input it in any way. The wrong stimulus will simply have no causal effect and the mechanism will remain idle. The m-context is not explicit in any way, nor is it comprising any kind of knowledge that underpins the inner workings of the mechanism. Thus, the m-context is not an explicit data structure, but a causal domain that can be taken in isolation. This domain underlies the mechanism's functional success conditions. But what does it buy us? We can see that by contrast with what would be an m-contextless machinery. Such a mechanism could grab the current auditory stimulus, hand it to a parser that would extract its structure and yield it to another mechanism, depending on what it finds. In this case, there would be no implicit m-context.

This amounts to a non-representational and non-computational version of the "just add another entry" approach we've just discussed. In order to avoid getting lost in a causal ocean, the mechanism must either assume that the m-context was established somewhere else or identify the m-context itself. Only then it can figure out how to handle the stimulus. This identification process can be quite complex and involve both bottom-up (from stimulus to m-context) and top-down (from a candidate m-context to stimulus) processes. The contrast with Wheeler's SPACs is clear, for in these mechanisms, m-context is always already there, at the point of triggering. Thus, Wheeler allows for a genuine explanatory role for m-context in accounts of the minds' inner workings: it is part of the functional description involved in some mechanisms. Consequently, we have something similar to what was achieved by the knowledge representation approach, and we can again make sense of what would it be for a sub personal mechanism to be "inside" a given m-context and have its inner workings relying on this.

Both Minsky's frames and Wheeler's SPACs enables an explanatory role for m-context in their respective frameworks. In order to do that, both work under the assumption that m-context boundaries can be somehow determined. But there's no such thing as a non-saturable data structure or a non-saturable isolated causal domain, so there can be no one-to-one mapping between m-context and human contexts. Furthermore, there is no clear principle to individuate neither frames nor SPACs, and no criterion under which we can decide where goes what (should cognitive knowledge about work colleagues be part of the restaurant mechanism?). In particular, this is probably the main reason why Minsky's approach could never achieve anything beyond specialized systems, in which the context is artificially constrained by the designer, allowing the system to handle a single task in a very specific set of circumstances. Wheeler's SPACs don't share the same historical importance, but they can be subject to the same reasoning. That doesn't mean the approaches are doomed, however, but only that there is a huge challenge ahead. A more detailed diagnostic will be provided in the upcoming sections. For now, I want to focus a bit on researchers who had already given up and believe that any appeal to m-context in this sense is hopeless. How can they navigate the huge ocean of cognitive knowledge without any appeal to m-contexts?

Take for instance Dreyfus (2007) and Bruineberg; Rietveld (2014). In their view, our minds are continually "tuned" to the environment in a more-or-less global sense. However, to claim that such tuning is global does not mean that every single mental mechanism involved is deeply integrated with one another. There is space for specialized or even encapsulated mechanisms. The point is just that the organization of these mechanisms need not rely on determined m-contexts structurally resembling human contexts in any way. This is also the reason why Gallagher (2007) thinks that m-contexts have no real explanatory work to do. In the view of these authors, although the human world can be organized in different contexts, such structuring has no role to play in explaining our ability to negotiate and navigate them. The mind is sensitive to lots of environmental cues, and behavior is to be accounted for without relying on their organization.

In this view, to enter a context is nothing like selecting an m-context to guide one's cognitive machinery. Instead, the whole system can be said to acquire a certain shade in a broad spectrum of possible tonalities. This sounds familiar, for that's how we characterized human contexts: contextual tonalities. But what exactly this buys us explanation-wise? *Prima facie*, by rejecting the need to compartmentalize and organize chunks of cognitive resources or to specify contexts in which one can embed special purpose couplings, we can avoid the need to deal with the non-saturable character of human contexts at the level of cognitive explanation. Contexts would be a methodological tool that we, as theorists or scientists, can apply to organize *our* work ("we study human behavior in restaurants"). Should we accept some version of the global state story and avoid any role for m-context in cognitive explanations, then? Though that would surely make Dreyfus happy, the situation is not hopeless for those who want to insist in an explanatory role for m-context. It's far from clear whether Dreyfus's or Bruineberg's and Rietveld's approach really buy us anything as far as context-sensitivity goes. The need for an account of the human capacity to cope with an open-ended set of non-saturable contexts is still there. Scientists adopting a global state story still owe us an explanation of how can the trajectory from one global state to the next be isomorphic to the structure of contexts we find in the world.

To provide the grounds for such a trajectory and to provide a theory of context determination, however, are very similar tasks, so even if one comes up with a proposal to handle the former, there's always a possibility that the solution can be formulated in terms of the latter. It is an open question whether there can be any solution-wise distinction that preempts such a move. This picture suggests that m-context issues cannot ground the selection of one explanatory framework over the other. Therefore, one cannot dismiss the problems brought about by the existence of non-saturable contexts just by pledging allegiance to a specific framework. Therefore, at least as far as m-context issues goes, no framework gets to handle commonsense and situation holism for free. As we'll see in the next section, this goes even deeper than one might think.

Before we move on, it is important to be clear about what motivates bringing up the possibility of m-contexts to our picture. Basically, it is because the debate around whether it makes sense to rely on m-contexts in explanations of cognitive capacities is historically mistaken as the debate about whether it makes sense to rely on representations. So we need to make a clear-cut distinction between issues regarding the (possible) reliance on m-contexts and issues regarding the (possible) reliance on representations in accounts of cognitive capacities and performances. As an example, when Dreyfus published his early criticism to AI, the challenge of accounting for context-sensitivity underpinned most of his arguments (Dreyfus, 1972). But at the time, as Dreyfus had no alternative framework to offer, he pointed the finger towards representations and claimed that cognitive science needs to avoid symbolic representations. As symbols were the only game in town, Dreyfus' concluded that we must shun any role for representational contents. In other words, he took the problem with the idea that m-contexts can be modeled as scripts or frames to be a problem with the idea that frame-systems and script-systems rely on representations.[22] The question of whether m-context has a role to play in cognition was conflated with the question of the nature and role of representations in cognition. However, we should be able to ask about the explanatory role of m-context in a given framework without worrying about its representational commitments, like we just did within Wheeler's framework.

1.4 How context sensitivity raises problems about relevance

Lets take stock of what we've already got. We saw that human intelligence is characterized by, among other things, two broad capacities called commonsense holism and situation holism. The first is the capacity to apply a huge backdrop of knowledge in real time as the interaction with the outside world develops (as inputs get in, if you like), and let this background guide our understanding of the current situation. Situation holism is the capacity to deal with two (or maybe more) of these understandings of the world at once. It'd be already impressive if we could do this while keeping these two interpretations apart, but we can also integrate them, and reason about how they interact with each other. I believe that both of these capacities are ubiquitous, that is, they underpin a large part of our cognition, even online, low-level processing, such as that involved in recognizing a tiger as a tiger. Furthermore,

[22] For a nice and elucidating exposition of Dreyfus's mistake, see Salay (2009).

the contexts involved in the structure of human world activities comprise an open-ended set of non-saturable situations whose boundaries are fuzzy and blurred. We have also seen how this trait of human world context raises an issue for cognitive explanations, and that this issue cannot be avoided by simply choosing one explanatory framework over another.

We're finally in the position to appreciate a deeper and broader issue that I'll call *relevance problem* (RP).[23] The conjunction of commonsense and situation holism with the non-saturable character of human world contexts raises serious problems when trying to handle relevance-sensitivity. The basic formulation of the problem is simple: given a goal and a set of resources, how do we select the relevant ones? Put this way, it may sound like a minor issue. There seems to be so many available options out there to deal with this question, that one can easily fail to see why the issue is worthy of our attention in the philosophy of cognitive science. As is typical with most of the philosophical problems, however, the original question seems banal and hardly worthy of serious research. It earns its keep as a concern by continuously resisting the naive initial approaches to cope with it. This, I think, is the case of RP in cognition. We start with a simple and vague question, and we realize that every time we advance a proposal to kick the problem through the door, it somehow makes its way back through the window.

What's the connection between the capacity to know what's relevant and what we've just seen about human intelligence and human world contexts? Given the initial question about relevance, any kind of cognitive system that must cope with it has only three broad strategies available: (1) always take for granted that every single available resource is relevant; (2) always take for granted that none of the available resources is relevant; (3) at any given case, take for granted that some of the available resources might be relevant.

For obvious reasons, we need not dwell on strategy (2). To say that a given system never considers any of its resources as relevant amounts to a denial of its status as cognitive. Only reflex answers like those produced by a knee that gets hit by a rubber hammer would be available for such a system.

In its turn, (1) is actually a way to shift the problem to someplace else. Unless we're talking about a system with an infinite amount of time and knowledge, we must assume that the available resources are manageable by the cognitive system.[24] This might be the case of simple enough biological systems which have only one or two broad strategies to deal with whatever situation they find.[25] Given a small enough amount of resources, one can always rely on everything one's got without worrying about time constraints. But how can a natural system possess just the right set of resources to cope with its environment? We know where to look

[23] Readers familiar with the *frame problem* (FP) may find this terminology a bit odd. Why not call such relevance issue "frame problem" as many authors do? Because I think these are different matters that call for different solutions. Though I have no terminological quarrel with those discussing relevance issues under the name "frame problem", I think the adopted terminology enables a clear-cut distinction and avoid common misunderstandings around the way the word "frame" is employed. The long answer will have to wait for the chapter 2.

[24] I'll leave God's epistemology for those interested in it.

[25] Again, I'm leaving aside whether such systems can be deemed cognitive. I'm just assuming that there can be cases of natural cognitive systems that might rely on always using every resource they've got.

for the answer at least since Darwin: the system's evolutionary history. Nice examples can also be found in AI systems. It is not a coincidence that in the history of artificial systems, this approach succeeds only for tasks executed in heavily (and artificially) constrained scenarios. Take for instance the so-called expert systems from the 1980s and the contemporary bleeding-edge applications of deep learning technology to create systems that achieve specific goals in efficient and robust ways. Such system's efficiency and accuracy rely heavily on the designer's capacity to constrain whatever it needs to consider when deciding what its behavioral output should be. The moral is that, if one wants to claim that us, *qua* cognitive systems, rely on the constant usage of everything we've got, one is only transforming the problem from that of selecting the appropriate subset of resources into that of finding a design solution that avoids this need. The resulting system would be one for which whatever the task, everything would be already in place.

It's a hard enough quest to find a design solution that can cope with a reasonable number of well-defined circumstances, but when considering the non-saturable character of human world contexts, it seems hopeless. It would be really odd to claim that one's knowledge about convolutional neural networks is *always* relevant in deciding what to feed a dog. Even if one's willing to bite such a bitter bullet, however, one would need to provide a design solution that shows how all this knowledge could not be an issue for resource-constrained systems. That is, she would have to show how the knowledge about convolutional neural networks can have the right kind of causal effect both when it is useless and when it is useful (as when one's thinking about machine learning). To show how such a distinction in causal effects could be done, however, is equivalent to show how a system can distinguish what is relevant from what is not. That's just the original problem. Therefore, even if our intuition could push us into thinking that commonsense and situation holism, along with non-saturable contexts, point towards (1), time and memory constraints would push us back.

It looks like we're stuck with (3), which is where RP really bites. If we want to explain how can our commonsense and situation holism deal with non-saturable contexts in a way that is compatible with our finite resources and the heavy time constraint that is imposed on us by their real time character, then we must understand how is it possible for us to keep track of what is relevant in every specific situation. The issue can be summarized like this: given a set Z of inference-like cognitive processes allowed by the constitution of a given system S — and assuming that Z is so big that it is physically implausible to expect that S discerns what's relevant through an exhaustive consideration of every single possible application — how can S select only the members of Z that are relevant for any given task inside any given context? Notice that Z need not be open-ended in order to preclude an exhaustive consideration. It just needs to be large enough so that the strategy of checking all possibilities is beyond the reach of the system.

As Haselager remarked, this is the problem of understanding how only the relevant part of our cognitive knowledge guides what we end up believing and doing (Haselager, 1997, p. 105). The most evident difficulty is that there seems to be no general theory of relevance nor anything akin to a set of general principles that might distinguish what is relevant and what is not in a contextless way. This is no surprise. Relevance is a relational property. A hammer cannot be said relevant *per se*. It is relevant for a given task in a given context. Relational properties cannot be easily (if at all) reduced to non-relational properties such as physical ones. Relevance is not a property intrinsic to members of Z: just like hammers, any of them might be relevant (or not) in one situation or another. As Fodor (1983) would say, Z is isotropic. But that means we're now in a very delicate situation: on the one hand, we need to account for relevance sensitivity in order to explain how we can negotiate human world contexts. On the other, we must rely on context to distinguish what is relevant and what is not. It seems we've got sucked in a vicious circle.

How hard can it be to get out of this circle? We can see that by assessing the way it was approached throughout the history of cognitive science. This will show us how easy it is to be misled into thinking that we have found a solution even if all we did was to reformulate it in a different framework.[26] That's the path we tread in the next session. For now, I just want to remark that, as Samuels (2010) has noticed, if this analysis is on the right track, the resulting picture is remarkably close to that on which Descartes has grounded his substance dualism: reason's flexibility and capacity to cope with an open-ended set of situations resists mechanistic accounts, and relevance sensitivity might be the one to blame. A science of mind that cannot account for it is dangerously incomplete, and that's the importance of trying to meet this challenge.

1.5 Why relevance problems are so hard

Throughout the years, cognitive science has developed many strategies that promised to handle RP. I think none of them gets there. The usual outcome is to find a non-generalizable strategy that ends up assuming commonsense rather than explaining it. This is typically the case among those working closely with AI research and using it as source of inspiration. The stance is reminiscent of cognitive revolution's first days in which AI was cognitive psychology's intelectual engine. In those times, AI was the primary source of computational models that cognitive psychologists could take as hypotheses about the inner workings of the mind. However, in AI, artificially constrained scenarios ("toy worlds") are typically already taken for granted when specifying the capacity to be modeled. Thus, just by specifying a target capacity in a certain way, one is already assuming the context in which it will be applied, which means the system relies on the designer's commonsense, and not on its own. Such a context will be then artificially saturated, for it is up to the designer to point out scenarios in which the system will not behave as expected and constraint the environment in a way that avoids them. Even if the AI designer manages to deal with a huge system who can cope with a large set of situations, her solution cannot be generalized in a non *ad-hoc* way, for it relies on lots of special-purpose artificial constraints she put there.

[26] A danger that surely lurks over the possible solution I'll present in this work as well.

But AI is not the only path to this error. Even those less computationally inclined and closer to biology can easily fall for it. The reason is closely related to a methodological difficulty that pervades psychology. Take for instance the McGurk effect (Mcgurk; Macdonald, 1976), according to which visual inputs may affect and override what we hear. If we hear someone saying "ba" (as in bark) but see her mouth moving in a way that is closer to what we would expect if she was saying "va" (as in vacation), we'll actually hear *va*, in a kind of auditory illusion. The mere specification of the effect (the conditions under which it happens) does not amount to a psychological explanation. The specification of an effect is psychology's *explananda*, not its *explanans* (Cummins, 1983, 2010a; Haugeland, 1998b). Their manifestation is something to be explained. The relevant point to our discussion is that the specification of capacities is no trivial matter. They're akin to what Marr (2010) dubbed the "computational problem" (a misleading name, for there is nothing essentially computational about it). Some specifications might be clear enough, such as the capacity to see depth in a visual input. Other capacities have boundaries that are much harder to find, such as the capacity to learn a second language. I have no intention to discuss this *qua* methodological problem, but I want to remark how easy it is to be blind about relevance sensitivity either as a necessary part of the capacity's specification or as something that the specification must take for granted.

To show how pernicious and ubiquitous this might be, I'll discuss how RP presents itself (and refuses to leave) in two largely different frameworks. The first is Fodor's classical view of the mind as a collection of modules and a global processing engine. The second is Wheeler's Heideggerian account of the mind. Fodor's view is connected to the classical cognitivist tradition and is broadly computational and representational. Wheeler's view trades on a typically non-computational and non-representational modeling medium: the Dynamic Systems Theory (DST) mathematical framework, usually associated with physics. What's interesting and elucidating is that, while the former assumes a destructive pessimism about RP and claims that there is no possible cognitive science of relevance sensitivity,[27] the latter tries to explicitly overcome it. Nonetheless, both end up in a similar place, with difficulties that are analogous in form, despite their huge differences.

1.5.1 On Fodor's pessimism

Fodor is well-known for taking two different claims to the extreme. First, a claim about the kind of representational vehicle that underpins our representational capacities (Fodor, 1980). In his view, attitude contents amounts to representational contents. His famous *dictum* makes it clear: to believe in p is to have a representation with the content p in the belief box. This is Fodor's way to say that to believe in p is to have some internal state that both represents p and has been given a functional role: causally grounding the behavioral outputs of the system in a way that can be epistemically assessed as taking p to be true. In the next chapters I'll have a lot to say about what the rejection of this claim buys us, but for now, lets stick with Fodor.

[27] Remember Fodor's "law of non-existence of cognitive science" in Fodor (1983).

The second claim for which Fodor is well-known regards the modularity of mind (Fodor, 1983). In this view, we can talk about at least two kinds of cognition: modular and central. Fodor's conception of modular cognition is very specific and his properties are long-familiar, so I'll be brief here and quickly list the most prominent characteristics that he attributes to modules. First, they're domain specific. Whatever a domain amounts to, the module is constrained by it and will refuse to handle anything else. Second, they're mandatory, that is, the presence of the adequate input will necessarily trigger the module into doing whatever it does. Third, they're cognitively impenetrable, that is, processes outside the module cannot interact (except maybe in very limited ways) with the module's inner workings. This is why the agent has no say in what the module outputs (think about perception: though you can refuse to accept that the wall ahead of you is indeed purple, you can't decide to see it differently). Finally, they're informationally encapsulated, in the sense that whatever content they have is proprietary, and whatever content other modules have will not be taken into consideration.

Sometimes Fodor is clear in saying that informational encapsulation is the most important characteristic he attributes to modules. However, even if this is not entirely true of his view, encapsulation is surely the most relevant property for our discussion. Perceptual input mechanisms are the usual examples of informationally encapsulated modules, and that's why one could point to the McGurk effect and take it to be evidence against Fodor's view. The effect, remember, is one in which visual input changes what we hear. If those processes were completely encapsulated from each other, how could that be possible? Though I believe Fodorians would indeed owe us an explanation for this kind of case, I don't think it would be too hard for them to produce one, nor that the lack of an immediate answer would be fatal to the encapsulation thesis. There are many possible stories. Maybe there's more than two modules involved in vision and audition, and one of these modules might rely on already worked out inputs from other, lower level, modules. Or there can be some kind of architectural failure that allows for this very specific effect. As a matter of fact, if we assume non-encapsulated modules, we'd have to provide an architectural explanation about why there's so little interference among them, which would put us in a position very similar to those working with a knowledge representation framework. Indeed, Minsky's frames and Schank's scripts, which we have met when discussing context, are both strategies that can be used to tame non encapsulated processing.

Fodor's insistence on encapsulation is a reflection of his pessimism about such taming strategies. This goes very deep. For him, being encapsulated and being an object of scientific study come together. Why is that so? Fodor regards non encapsulated systems as central systems, the ones responsible for non-modular cognition. Such systems have access to each and every piece of information available in the system as a whole, and his main job is to integrate the outputs of all the modules in some kind of global or system-wide processing. Fodor's most common example is belief fixation: given that the system already has a set of beliefs Σ and receives a new input, what should it end up believing as a result? That is, what should change in Σ? We have already seen how difficult it is to answer that question. A satisfactory answer must rely on a good account of commonsense and situation holism, which in its turn must rely on a good account of relevance sensitivity. The unavoidable question thus becomes: how does only the relevant portion of Σ affects what the system ends up believing?

Fodor's point is that there is no possible scientific account for central cognition. If that's true, then there can be no scientific account of human intelligence. To ground such an extreme claim, Fodor characterizes central cognition as both *isotropic* and *Quinean.* To say that a system like Σ is isotropic means that all of its elements are potentially relevant for the processing of any given input. In belief fixation, this means that every single belief is potentially relevant in processing the input, and thus no one can be ruled out as irrelevant *a priori.* To say that a system is Quinean means that, even if we find some non-arbitrary way to distinguish relevant from irrelevant subsets in central systems and avoid the need to make an exhaustive consideration of Σ for every single cognitive task, the solution would not be generalizable. That's because the current circumstances (the conjunction of the current state of affairs in the world and the contents of Σ) can change what is relevant and what is not. That's Fodor's path towards the issue of handling the non-saturable nature of human world contexts. He thinks that there is no way to tame central cognition, and that approaches like Minsky's frames are hopeless. The best one could get by treading that path is an adequate subset of what's relevant for a very specific and concrete circumstance. This is a far cry from where we've been expecting to get.

As hard a problem that is, it's not yet clear why would Fodor claim that central cognition is beyond scientific reach. As a first clue, we have to remember that Fodor was a classical computationalist: to cognize *is* to compute. In this view, we can study and understand biological cognition only by studying and understanding computational models. However, there is no possible computational model of any system that's isotropic and Quinean, *ergo* there is no possible scientific account of central cognition. On the other hand, it is possible to create computational models of encapsulated modules, and thus we can have a cognitive science of such modules. Computationalism is therefore incompatible with non-modular systems. But as computationalism is the only game in town, we're out of luck.

I'm not committed to a broadly computationalist approach, but I find instructive to assess whether Fodor's worries are on the right track.[28] Assume computationalism is all we have. Does it mean we must accept such a pessimist conclusion and narrow down our scientific enterprises to modular cognition? I don't think so. In the Fodorian picture, belief revision is described (explicitly) through an analogy with theory acceptance in science. The process through which our cognitive apparatus decides how Σ must be reviewed in face of a new fact is akin to the scientific process of reviewing its whole body of knowledge as a function of some new fact. If scientific theory review is both isotropic and Quinean (as Fodor assumes), then so is belief revision. But that's simply not true, for there are other possibilities.

Scientific theory acceptance can also be conceived in a broadly sparse fashion. We can have groups of scientists or scientific enterprises in general that are completely ignorant of

[28] I do believe computational mechanisms have an important role to play in explaining biological cognition but, *pace* Fodor and classic cognitivism, I don't think that computers can tell whole story.

what the others are doing. Science's isotropic and Quinean consensus might be a product of the ocasional clash and competition between these relatively independent enterprises. In usual circumstances, individual scientists or small groups simply won't regard anything outside the boundaries of their objects of study or research tradition. In this view, properties like being isotropic and being Quinean can be said of the whole body of scientific knowledge, but they emerge out of the interaction of many local enterprises and not out of a global consideration from every single scientist or research group. I don't know whether this is an accurate (yet very sketchy) picture of how science consensus comes to be, but it doesn't matter much, for my only point here is to show that Fodor cannot take scientific reasoning as grounds for such an extreme claim about the impossibility of modeling central cognition. What Fodor thinks about cognition is constrained by what he thinks about science, but his conception of scientific reasoning might well be wrong. We can (and I think we must) distinguish Fodor's reasons for his pessimistic claims about what can or cannot be an object of study for cognitive science, from the grounds he offers to show how deep and hard is RP. The former relies on a questionable conception of scientific reasoning and a firm belief that computationalism is all we have, while the latter is a problem even for those who disagree with Fodor about science and who do not endorse any form of computationalism. [29]

Once we take Fodor's computationalism and conception of scientific reasoning out of the picture, his position can be summarized as follows: encapsulation is the only way to avoid RP. Why is that? Because given the specialized and domain-specific nature of the tasks that modules deal with, they can get by without wondering about relevance. They can apply strategies such as using every resource they have, since they don't have many. In other words (closer to mine), modules need not rely on commonsense holism or situation holism, and they need not worry about the non-saturable nature of human contexts either. Of course, to say that they need not worry is different from claiming that they managed to solve the problem. What I mean is that they don't even try. Cognitive capacities that rely on relevance sensitivity are simply not their business. In the resulting picture, it remains a mystery how can we account for central cognition, but not because of a commitment to cognitivism or computationalism. Rejecting these simply won't help.[30] The problem is that the strategy available to modules (given an input, they can select an adequate subset of all information they have through an exhaustive consideration) is not an option for central cognition. Remember our discussion about contexts: even if we bite the bullet and claim that belief revision *does* encompass the consideration of all of our cognitive resources at every single step of our continuous interaction with the world, we'd still owe an explanation of why any given resource (a belief, for instance) sometimes has a causal role that can be epistemically assessed as that of being relevant for the current task, and sometimes it has not. The question of how can everything we know bear on what we end up believing or doing is shifted into the question of how can the

[29] This distinction will play an important role in the forthcoming assessment of Dan Sperber and Deirdre Wilson's approach to relevance.

[30] We'll see details on how can a non-computational approach fall for similar reasons in the next section.

relevant part of what we know bear on what we end up believing or doing.[31]

At this point, computationalists in general, as well as those closer to AI research might be tempted to regard this as a problem about computational power. Once we have a powerful enough computer, be it a quantum computer or whatever, the consider-everything strategy will prove itself viable. For those with that inclination, I just want to remark that (1) this is not true; and (2) even if it were true, it would amount to a brute force approach that might be useful for some AI systems, but it won't help cognitive science to understand how *we* do it with the amount of resources we have. To ground my claims, I'm going to use a calculation made by Cherniak (1990): assume Σ has only 138 beliefs, and we need to check for the consistency of the conjunction of these 138 sentences by means of a truth table. How long would it take? Depends on how fast our computer is, of course. Let the time spent computing each line of the truth table be the same a ray of light (traveling at 299.726 km/sec) would take to traverse a proton (whose diameter is 10^{13}), which is something around $2{,}9.10^{-23}$ seconds. Such a computer would need more than 20 billion years (remember, the table has 2^{138} lines), which is roughly the estimated time from the big bang to the present. Such a calculation might be fun, but will probably raise lots of reasonable questions: why a truth table? There can be better ways to structure the data and better algorithms to apply. True, but that's exactly what leads to the point (2): to find a more efficient algorithm amounts to find a way to go on without the exhaustive consideration of every possibility. It's primarily about what we do, not how fast we do it. Talk about computational power is simply beside the point. AI *qua* business enterprise might benefit a lot from the increasing availability of more and more computational power but this won't necessarily matter much for the cognitive psychologist who's worried about how can *we* do it.[32]

Under these circumstances, there are two possible roads worth considering. The first is to reject Fodor's idea that we can only get rid of RP through encapsulated modules and find a way to tame central cognition. The second is to embrace the idea and take it to the next level by rejecting the existence of central cognition: we're modular all the way up. What both approaches have in common is that they consider the distinction between modular and central cognition unrealistic. Let's take a quick look at both.

1.5.1.1 Can we dodge Fodor through heuristics?

Let's start with the first road. Some authors, such as Carruthers (2003) and Samuels (2010)[33] have claimed, *contra* Fodor, that encapsulation is not necessary for taming central cognition. In their view, the key to ensure the computational tractability of centralized processing is to use heuristics. A heuristic is a kind of rule of thumb that allows the system to get by in a

[31] Note that guarantees are beside the point. We can fail to see what's relevant. Indeed, we fail more than we'd like to admit. Still, our performance goes way above chance. As Fodor would say: contemplate our not all being dead.

[32] Indeed, the contemporary hype with deep learning models is partly due to the increasing availability of GPU's processing power (units dedicated to graphical processing).

[33] See also Samuels (2005).

fallible yet reliable way. This kind of rule is the reason we can play chess with our computers. Deciding what's the best next move in a chess game is computationally unfeasible. However, by adopting heuristic strategies, such as "in the first 10 moves of the game, one must focus on dominating the central area of the board", the system can significantly reduce the amount of possibilities, for it need not find the best available move, but only the one that increases its dominance in a well-defined subset of the board. Notice that, by following this rule, the system may become blind to some unusual opportunity that happens to take place. The opponent might make a move that renders her queen vulnerable, but if the system focuses solely on dominating the center of the board, it ends up failing to realize that. Ain't that exactly the kind of mistake that happen to us all the time? Sure. This balance between reliability and fallibility is what makes heuristics attractive. If this is on the right track, central systems may be able to cope with the prohibitive amount of information that characterizes them, for they're always handling just the subset selected by the currently applied heuristic.

Unfortunately, if left on their own, heuristics fall short of what we need. They do provide computational tractability, but only because they shift the burden to how information is organized in the system. To see this, consider what would be the outcome of a heuristic rule that is completely insensitive to the underlying organization of the information it is working with. Here's a quick and dirty example guaranteed to render any problem computationally tractable: whatever the problem, the processing is going to consider only up to five inferences and ignore the rest. The obvious question would be: which five? Is the chess system going to consider the first five possible moves of the leftmost piece, or only the first possible move of the five leftmost pieces, or...? The answer relies on how the information grounding the inferences is organized. Consequently, if the organization does not suit the system's purposes, the behavioral outcome may be erratic.

Using organization to cope with huge amounts of information is a strategy we have already seen when discussing the role of m-contexts in cognitive explanations. The same idea can now be put in different words: we can tame central cognition by compartmentalizing and organizing the available information. The most well-known approach of this kind is Minsky's frames: the crucial idea is that one can organize the information concerning different domains (or situations, or kinds of situations, and so on) in different data structures and link the possible interactions among them. It is worth to let Minsky himself introduce his idea:

> A frame is a data-structure for representing a stereotyped situation, like being in a certain kind of living room, or going to a child's birthday party. Attached to each frame are several kinds of information. Some of this information is about how to use the frame. Some is about what one can expect to happen next. Some is about what to do if these expectations are not confirmed. We can think of a frame as a network of nodes and relations. The "top levels" of a frame are fixed, and represent things that are always true about the supposed situation. The lower levels have many terminals—"slots" that must be filled by specific instances or data. Each terminal can specify conditions its assignments must meet. (The assignments themselves are usually smaller "sub-frames.") Simple conditions are specified by markers that might require a terminal assignment to be a person, an object of sufficient value, or

> a pointer to a sub-frame of a certain type. More complex conditions can specify relations among the things assigned to several terminals. Collections of related frames are linked together into frame-systems. The effects of important actions are mirrored by transformations between the frames of a system. These are used to make certain kinds of calculations economical, to represent changes of emphasis and attention, and to account for the effectiveness of "imagery." (Minsky, 1997, p. 156)

Through Minsky's description, it is easy to see how he thinks that frames might be able to tame central cognition. By selecting which information goes in which data-structure, and by mapping the possible interconnections among them, central cognition need not get lost in idle processing about completely irrelevant and unrelated knowledge. As we have already seen, frames are a way to account for the role of m-context in sub-personal mechanisms. Unfortunately, they don't work, and we already know why: frames are a way to specify contexts, but human contexts cannot be specified in full because of their non-saturable character. If one insists in using frame-systems, one quickly ends up in a non ending spiral of increasingly complex frame structures for dealing with increasingly refined and specific contexts (from *restaurant* to *restaurant-with-date* to *restaurant-with-date-without-work-colleagues*, and so on).

In the end, we get either a set of frame structures describing artificially (and perhaps arbitrarily) saturated contexts (as we do in chess engines), or a system that fails to tame central cognition. On the former case, we fail to account for the target phenomenon, which is the non-saturable nature of contexts. If we take a frame to be the ultimate specification of any context, the system won't be able to cope reliably with any new element that presents itself in that context (a fly, an ex-boyfriend, a blackout, and so on), for it will simply be ignored. On the latter case, the frame system would have only a partial specification of each context, leaving the rest to be determined in real time. The frame would not be a specification of the context, but a kind of stereotype the system can use as a starting point to deal with the real thing. But that begs the question on how we can determine it in real time. Frames were supposed to explain how we do this kind of thing. What we have thus is not an account of how central cognition can be tamed, but one that assumes it has already been tamed somehow.

If frames can't tell the whole story, perhaps they can at least provide a good starting point for the application of heuristics? That is, maybe frames could account for what's stable (or typical) in the human world, while heuristics allow the system to go beyond that and handle whatever is specific to the current circumstances. Thus, if the restaurant frame has no information about the appropriate course of action to take when the waiter is rude, one can rely on heuristics to find out the answer in some other frame. Unfortunately, as it stands this suggestion is hopeless. Heuristics cannot alleviate the burden over information organization because they rely on it themselves. Even this search for answers somewhere else must rely on the world being organized in some way. Therefore, situations in which the world (as modeled in the frame system) is organized in non-usual ways will remain out of reach.

Consider another example: say you have to find a book in a huge library just by looking at the shelves.[34] In the worst case you'd have to check through each and every book until you find the one you want. But here's a nice heuristic strategy that might help to avoid that: take the first letter of the author's last name and use it as a guide. If it is a "Z", then you can save a lot of time by starting with the shelves at the end of the corridor. If it is a "K", you can start the search around the middle of the way. In both cases, if you see books whose author's last name starts with a letter that is alphabetically closer to the one you're looking for, you can use it as a clue for where you should go next. A smart strategy indeed, but as you probably already realized, the effectiveness of this strategy depends on how the books are organized. If the books are ordered by color, it simply won't work and then (assuming you have no information about the color of the book) you'll have to look one by one.

This example makes salient that heuristic strategies must rely on certain assumptions about how the world is organized. In frame systems, to provide such assumptions is the main role of frame data structures. In the absence of the necessary information, not only heuristics are blind, but the system has no clue to choose the appropriate heuristic among the set of available ones. If the system has no idea how the books are organized, he can't choose between, say, looking for the author's last name or looking the color of the cover (assuming it knows both strategies). Well, some might ask: couldn't we try both? Couldn't we try all strategies we know, assess the results and improve whenever possible? In a sense, I believe that's exactly what we do, but frame systems are unable to explain how. This is only viable if we have a small enough set of sufficiently broad strategies. In frame systems, this is not possible unless we artificially restrain the set of possible situations the system can cope with. A heuristic like "do whatever is urgent for survival" might be adequate for life or death situations, but should it be applied if your favorite dish is unavailable at a restaurant?

Furthermore, as we have seen when discussing human contexts, the addition of a new feature to a new situation does not amount to a simple addition of new information in some frame slot. New elements, like new cues in a crime investigation, have the potential to change the way the system should regard everything it has already considered. We have thus a vicious circle: new inputs might have an effect on the way we organize information. They can change causal relations previously assumed and can drastically change the weight that some previous feature bears on the current circumstances. Remember the example of the kids playing pretend. Is the voice of their parents calling from the other room to be taken as a sign that it's time to stop or could it be integrated into the play somehow, thus allowing the kid to approach his dad by saying "who dares to call the King"? A lot of different inputs might change the behavioral output in one way or another. The only resource frame systems have to account for these many ways is by explicitly specifying them. Thus, if we want to account for how new inputs can reorganize a frame structure, we need a kind of second-order frame structure describing how it is done. But of course the frames that describe how the first-order frames can be affected may themselves be affected by some input, which leads us to the need for a third-order frame

[34] This example was adapted from Chow (2013).

system. This is a clear case of a vicious regress. Some like Dreyfus (1972) saw this threat right after frame systems were first presented. In a nutshell: we can't account for how to cope with non-saturable contexts by postulating second-order non-saturable contexts to lead the use of the first order ones. But if we use heuristics within a frame system, that's all we can do.[35]

1.5.1.2 Can we dodge Fodor through massive modularity?

What about the second road? The core idea is to take the modularity thesis to an extreme and go modular all the way up. In this view, the mind is essentially a product of the complex interaction between many encapsulated modules. The mind is massively modular, as some would say. Consequently, there is no such thing as central cognition to be tamed. Many researchers connected to evolutionary psychology, such as Pinker (1997) and Sperber (2005), chose to tread this path. There is extensive literature about this view and I cannot do justice to it: even a birds-eye view would take too much space. My discussion will thus be restricted to if and how a massively modular mind is able to cope with relevance sensitivity in the characteristically human way, even when handling non-saturable contexts.

As far as RP goes, Fodor's point is that encapsulated modules need not worry about it. A well-functioning module just need to cope with a specific kind of task by using a proprietary, predefined and statically organized information corpus. If the module's performance is good enough to be adaptive, then it's good to go. However, these are also the properties that seems to preclude the appeal to modules in any task that relies on information integration. A facial recognition encapsulated module that has no access to geographical information will be of no help in guessing where someone comes from. This kind of informational boundary led many to think that accounting for relevance sensitivity in massively modular systems is hopeless, as they seem to allow for near-zero flexibility and context-sensitivity. Though I agree that massively modular systems won't do, that's not because they have no flexibility at all. There is space for some flexibility if we consider that modules can eventually preempt each other (as when a predator-detection module takes over) and that modules can get triggered by complex input clusters (a visual and an auditory simultaneous input, for instance), but of course this is a far cry from what we need.

According to Sperber (2005), modular systems can be much more flexible than that, and are able to account for context-sensitivity. I'll quickly present what I take to be his core claims about this because it will help me point out exactly why I think massive modularity falls short of what we need (just like frame systems). Sperber claims that the kind of flexibility and context-sensitivity that we exhibit is due to a general organismic tendency to improve efficiency at all times. His background for this claim is the theory of Relevance he developed with Deirdre Wilson (Sperber; Wilson, 1995). There, they present a technical notion of relevance that I'll call *s-relevance*, in order to avoid confusion with the more intuitive sense we've

[35] At this point, one can say that the problem with frame systems is that they are computational, and hence they must regard relevance as a computational problem. This will led to the road where some try to avoid RP by avoiding its computational guise. In the next section I'll try to show why the effort is not fruitful.

been using the term so far. The core idea is that s-relevance is a property of inputs that can guide the effort and the amount of resources (such as memory retrieving) dedicated by the cognitive system in attending it. If there is no central cognition to assess the input and control the allocation of such resources, this is achieved through competition: the cognitive economy must be so that the most s-relevant input is always (or usually) more likely to get most of the available resources.

In order to see how that could work (and ultimately why it falls short of what's needed), we have to understand a bit about what it means to say that a given input is s-relevant: the measure is done in terms of a trade off between cognitive effort and cognitive benefit. A cognitive benefit is an effect that contributes for the general efficiency of the encompassing organism. In organisms like us, the idea is that the more we know about the world, the better (i.e. the more efficiently) we can cope with it, and thus we can grasp how cognitively beneficial some input is by measuring how many new inferences it allows us to do in a given set of circumstances. Remember, however, that s-relevance is about a trade off. It takes the system time and effort to cope with the benefits of any given input, so the most s-relevant input is not just the one with more cognitive benefits, but the one with the greatest set of benefits at the lowest cost. If any two inputs allow for, say, three new inferences, then one that requires less cognitive effort will be regarded as the most s-relevant. [36]

A question quickly emerges: how can any cognitive system be sensitive to s-relevance? At any given time, the s-relevance of an input depends on the current circumstances and the current cognitive states (which encompasses the organisms current needs and expectations). Given such reliance, it seems that a system can only know the s-relevance of some input *after* a lot of processing. But if that were the case, then the whole s-relevance framework would be self-defeating. What really matters then is not sensitivity to s-relevance, but to expected s-relevance. If the system is able to calculate the expected size of cognitive benefits, it can then allocate the right amount of resources for the input at the point of triggering. In the resulting picture, cognitive efficiency is partly measured by how accurately can a system make well-informed guesses about the s-relevance of an input.

But the worry about how can an input's expected s-relevance be calculated remains. After all, if the calculation is not efficient, then the whole project falls apart. In our previous discussion about frame systems we have faced some difficulties that might reappear here. Are we going to use heuristics? Which ones? Perhaps it's possible to measure the number of inferences that a given input allows and then store it, so we can use that as a shortcut when processing the same input in the future. But this strategy is ignorant of the particular importance that any given cognitive benefit might have in the present circumstances, i.e. given the current needs and goals of the encompassing system. After all, s-relevance, just like relevance, is a relational property. The s-relevance of any input is a function of the current circumstances

[36] This is of course a very simplified description of a single element in a very rich and complex framework. But hopefully this is a fair enough gloss of Sperber's position, and I believe none of the missing details would have impact over the main point being presented.

and the current psychological states. By following this path, one might end up formulating an entire frame system (or something similar) designed to calculate expected s-relevance in a way that fulfills the goal of contributing to the system's efficiency in general. Sperber is well aware of that risk, and his suggestion is that the threat might be mitigated in a non-cognitive way: *"(...) it is not at all obvious that the brain should calculate the size of cognitive effects. There may be physiological indicators of the size of cognitive effects in the form of patterns of chemical or electrical activity (...)"* (Sperber, 2005, p. 65).

What's remarkable about this suggestion, whether plausible or not, is how much of a *detour* from the classical program of cognitive science it is. Sperber clearly endorses some version of computationalism, but he's willing to give up on the claim that the mind's sensitivity to s-relevance is achieved computationally. Of course, he does not deny that some of it might be computed, but his picture is one in which that would be an exception, and not the rule. Given the difficulties involved, I suspect this is no accident. Sperber acknowledges a clear role for m-context in cognition, but not a computational one.[37] This has important consequences. Remember that, in order to deny central cognition, we had to give up on the idea of a centralized ruler and put a decentralized competition for resources in its place. Thus, we have to account for the resource competition in physiological terms. It is worth to check Sperber remarks about that at some length:

> When an input meets the input condition of a given modular procedure, this gives this procedure some initial level of activation. Input-activated procedures are in competition for the energy resources that would allow them to follow their full course. What determines which of the procedures in competition get sufficient resources to trigger their full operation is the dynamics of their activation. These dynamics depend both on the prior degree of mobilization of a modular procedure and on the activation that propagates from other active modules. It is also quite conceivable that the mobilization of some procedures has inhibitory effects on some other procedures. The relevance-theoretic claim is that, at every instant, these dynamics of activation provide rough physiological indicators of expected relevance. The flow of energy in the system is locally regulated by these indicators. As a result, those input-procedure combinations that have the greatest expected relevance are the more likely ones to receive sufficient energy to follow their course. (Sperber, 2005, p. 68)

Metaphorically, we can take such physiological markers to comprise a complex road-like structure that routes inter-module interaction. Such routes may be established both genetically and developmentally. Thus, even if the interaction between different modules occurs only through well-defined interfaces (i.e. only through accepted inputs and the resulting outputs, with its inner workings remaining out of touch), the resulting picture allows for a fairly complex interaction. We can develop the metaphor a bit further by imagining that some routes may have "obstacles" (that is, something inhibitory) or being the kind of route that allows for

[37] There's room for some confusion here. In Sperber; Wilson (1995), the authors use the word "context" in a very specific and technical sense that is distinct from the conception of m-context previously discussed. Its explanatory role, however, is quite alike: Sperber thinks that the activation patterns of mental modules have a role to play in explaining how we achieve sensitivity to s-relevance in sub-personal mechanisms.

faster speeds (or average speed, but demanding less resources) than others. Thus, certain inputs might fare better on certain inter-module routes than others. This metaphor helps us in understanding how physiological markers can guide the process throughout different modules in a way that is sensitive to expected relevance. That in virtue of which a given input is taken to be more or less s-relevant is a feature of the available routes. The system can thus allow for inferences to be made or preempt others by relying on the expected s-relevant without processing it.

Sperber's approach is fruitful and shows that massive modularity can account for a level of flexibility that could be easily left unnoticed. Given our current interests, however, the pressing question is this: is it enough to handle RP? Or could it at least throw some light on a possible path towards solving it? In order to be fair with Sperber, we should first ask ourselves whether RP is among the targets his account aims at. Given the comprehensiveness of the s-relevance theory, I think it should be. I'm not so sure whether Sperber and Wilson would agree, though. In a review of their book *Relevance*, Chiappe; Kukla (1996) have pointed out that the s-relevance theory falls short of a solution to RP. The authors took Fodor's construal of the issue, synthesized as "Hamlet's problem":

> If, for example, you undertake to consider a nonarbitrary sample of the available and relevant evidence before you opt for a belief, you have the problem of when the evidence you have looked at is enough. You have, that is to say, Hamlet's problem: when to stop thinking. (Fodor, 1987, p. 140)

That's just the previously discussed Fodorian perspective according to which belief revision processes are isotropic, Quinean and akin to scientific theory acceptance. Chiappe and Kukla went on to argue that s-relevance is unable to explain how the mind can solve Hamlet's problem, since it doesn't allow for a theory of central cognition, and nothing short of it will do. But this strategy left their flanks vulnerable to Sperber and Wilson's reply which is, in a nutshell, to claim that the mind doesn't work that way (Sperber; Wilson, 1996). It is neither Quinean nor isotropic. Rationality, *pace* Fodor, does not require that every relevant evidence be taken into account, only some of it. Sperber and Wilson explicitly connect the deep relevance issue that Fodor describes to Fodor's use of scientific reasoning as a model of cognitive processing. Consequently, Sperber denies that cognition is isotropic and Quinean, and claims that his account in terms of competition for resources is enough: *"In such conditions, Fodor's Hamlet problem has a simple in-principle answer. Let the process with greater expected relevance win."* (Sperber; Wilson, 1996, p. 532)

Were Fodor's account of the mind a necessary condition for the emergence of RP, Sperber's reply would be enough to settle the matter. But it's not. As we have seen when previously discussing Fodor's grip on the issue, RP emerges even if we take belief fixation processes to be less than global or to rise out of narrower processes. Therefore, those willing to claim that RP (as construed here) is not an issue for frameworks relying on s-relevance need more than Sperber's reply to Chiappe and Kukla.[38] If s-relevance is unable to handle RP, how does Sperber's suggestion regarding context-sensitivity fare against it? Unfortunately, just like frame systems, Sperber's physiological markers fall short of what we need: they can handle saturable contexts, but non-saturable ones are still out of reach. Consequently, if someone tries to apply Sperber's approach to RP, she will get nothing but a non-computational guise of the issue. Let's unpack this claim.

Remember that frame systems were able to account only for a somewhat inflexible background comprising a previously familiar set of artificially saturable contexts. Whenever such systems need to handle something beyond the typical flow stored in a frame (e.g. a new permutation of some situation), the frame stops being useful, and the system gets lost. In the same vein, physiological markers might enable the system to handle situations it is already familiar with, but won't help it in coping with unknown permutations. They can surely hiccup whenever something unexpected breaks the typical flow, but it can't tell the system what to do, for either it has no clue, or it has too many to choose from. Again, the addition of a new contextual feature can result in a broad revision of the present context. That's why a *restaurant* context might led one to very different behavioral outputs in contrast with a *restaurant-with-a-date* context.

The friend of massive modularity has two options here: either she dramatically increases the number of modules and potentially duplicates a lot of information among them in order to cope with these variations, or she increases the number of possible routes among the modules. The former is simply not realistic, for there would have to be uncountably many modules with redundant information and no place to stop the need for new ones, thanks to the non-saturable character of human contexts. The remaining option is to try to account for the many levels of s-relevance that a single input can have by increasing the number of routes among the modules. There could be a less demanding route and a more demanding one, a slightly less demanding one, a slightly slightly less demanding one, and so on, in an open-ended fashion.

But now we have RP all over again. In such a scenario, it is likely that we would need routes from almost every module to almost every other, as well as uncountably many physiological markers to account for the many shades of expected s-relevance attributable to an input. But if that's the case, then we're back to an unrealistic scenario akin to the formulation of a frame system that explicitly establishes all the possible informational routes. Expected s-relevance cannot modulate the flow anymore and something else is needed, for now the system has uncountably many possible s-relevance expectations to choose from. The only possible candidate I see for this something else is the organization of module routings in a way that is structurally similar to the contexts of the human world, which is, remember, exactly what frame systems try to do from the start, and also the reason why their limitations and difficulties will haunt Sperber's approach as well. In the end, even if it's true that Sperber's suggestions allow for a lot of flexibility in massively modular systems, this augmented flexibility is not nearly enough

[38] I'm not sure whether Sperber or Wilson would be willing to make this claim, but I'm inclined to guess they wouldn't. In any case, it's elucidating to see what would be the problem with it.

to account for the relevance sensitivity that is present in our target phenomenon.

1.5.1.3 Taking stock

If I was mildly successful in my presentation of both roads, it should be clear that they fall short of handling RP. Whether we're talking about data structures exploited by a central processor or physiological markers exploited by many encapsulated modules, the remaining question is the same: if and how these resources can be organized in a way that allows the system to handle non-saturable contexts. This result already suggests that RP is not computational in nature. We did try to follow Sperber's non-computational approach and ended up hitting a wall with a picture of Fodor's smile painted on it. But it remains to be seen whether that's an issue with Sperber's account or whether it would plague every non-computational approach. In what follows, I'll make the case that even if we deny space for modules, frames, computations and representational contents altogether, the same kind of trouble reappears.[39] We have to be careful, though. Getting rid of these cognitive resources often leads to adopting approaches that imply alternative explanatory frameworks. For instance, cognition as computational is frequently rejected in virtue of an understanding of cognition as situated and better modeled as a dynamical system.[40] In such frameworks, however, there is an additional danger RP-wise: they seem to be so distant from the kind of cognitivism we've been discussing so far, that we end up believing that the issue is gone for free. Thus, it is important to understand whether and how the problem arises there as well. Otherwise, we might not be able to see the difference between avoiding it and giving up on it.

1.5.2 On simplifying things by bringing Heidegger

For a while, the mind as a device computing over representations of the world was deemed as cognitive science's only safe place. No wonder, for the alternatives were the behavioristic distortion of mental phenomena and the methodological issues of introspectionism. But after decades in power, the limitations of computational approaches started to show and (worst) refused to leave. This led many to wonder whether there could be other safe places. This place was found in the idea that cognition is *situated.* Rather than conceive the mind as a device registering the world in a proprietary format to compute over, we should take the system-environment coupling as the basic unity of analysis. Some well-known proponents of this kind of approach are Brooks (1991), Varela; Rosch; Thompson (1991), Van Gelder (1995) and others. Nowadays, there's a whole ecosystem of situated, non-computational and non-

[39] There's an underlying discussion about the nature of computation and its relation to representation. Some like Fodor (1981) and Haugeland (1998c) advocate the semantic conception of computation, according to which there's no computation without representation. Others like Piccinini (2015) advocate for a different, mechanistic approach which allow computation over non-symbolic elements. Given my sympathy for Piccinini's account, I've said that we can deny both computation and representations and still be haunted by RP, but friends of the semantic conception need not worry, for that will make no difference in the foregoing discussion.

[40] As an example of this point, see Chemero (2009).

representational frameworks, and we usually refer to it as 4EA (embodied, embedded, extended, enactive, afective) cognition. For instance, there are ecological approaches - inspired in the early work of Gibson (1979) - such as the one from Chemero (2009) and varied enactive approaches like that of Noë (2004) or Hutto; Myin (2013). Of course, not everybody thinks that situated cognition implies the rejection of computational mechanisms and representational states or processes. Examples of compatibilist approaches are that of Clark (1997a) and Piccinini (2021). Within the 4EA ecosystem, the only such framework in which RP was explicitly addressed as an issue was the so called *Heideggerian cognitive science.* That's why the following discussion will focus on it.

Though some might find odd to bring Heidegger to the table, this is not really breaking news. In a historical sense, the whole 4EA research tradition builds on ideas that can be traced back to the phenomenological work of Heidegger and Merleau-Ponty.[41] That's the case of Hubert Dreyfus. A large part of his early critique to AI (Dreyfus, 1972) is underpinned by his view on Heidegger.[42] Dreyfus' book dedicates a large portion to sketch what would be a set of Heideggerian grounds for cognitive science, and those ideas share a lot of what grounds situated cognition. However, when the book was first published, his suggestions were largely ignored. The likely two main reasons are: first, most attention was given to the chapters where Dreyfus draws an acute (and sometimes aggressive) criticism of classical symbolic AI - famously dubbed *Good old fashioned artificial intelligence* or GOFAI by Haugeland (1985). Second, because Dreyfus' sketchy Heideggerian remarks produces no real alternative framework for cognitive science. He provides some phenomenological descriptions of human capacities related to intelligence and argue that GOFAI won't be able to model them. But what's needed was to outline a new research paradigm, and Dreyfus was never able to do that on its own.

More recently, however, Wheeler (2005) engaged in the project of formulating a scientific framework directly inspired by Heidegger's work in a sense that goes beyond situated cognition. His project's goal is, in a nutshell, to naturalize Heidegger's *Dasein.* Wheeler's basic approach involves moves like this: first he takes some Heideggerian element, such as the different ways in which we can render the world intelligible (ready-to-hand, ready-at-hand, presence-at-hand...) and then he aims at describing what could be their sub-personal marks, that is, what kind of mechanism could account for them.[43] Heidegger's scholars usually find this move to be heretic or unfaithful to Heidegger's original project, which is why two warnings are in place: first, it is not my aim to dwell on exegetical questions about what would be an

[41] Remember that Andy Clark has an important and influential book named Being There (Clark, 1997a), which is a literal translation of Heidegger's *Dasein.* Furthermore, Clark explicitly points to Heiddeger and Merleu-Ponty's ideas as providing early inspiration for the views he advances there. The same influence can be seen in Haugeland (1998d).

[42] In Dreyfus's case, his main influences are Heidegger (2012), Merleau-Ponty (2011) and Todes (2001).

[43] Readers not familiar with Heidegger's work must notice that this correlation between personal level phenomenological descriptions and sub-personal mechanisms is nothing like Titchener's attempt to get insights from introspection. Phenomenology is about describing the general traits of the human experience, not about describing particular subjective experiences.

appropriate reading of Heidegger's work.[44] Second, I'm not interested in questions about what would it take for a framework to be "genuinely" Heideggerian in spirit or even if such thing is possible without rampantly distorting his work. My discussion will focus solely on how well (or bad) this approach copes with RP. I'll start by presenting the core points of a debate between Wheeler (2012), Dreyfus (2007) and Rietveld (2012) about how should a Heideggerian framework handle it. After that, I'll argue that the upshot of the proposed framework amounts to neither a solution nor a dissolution of the targeted problem. RP is still there, pretty much untouched.

Wheeler, Dreyfus and Rietveld are all deeply committed to dynamic models built with the mathematical framework of the Dynamic Systems Theory (DST). This is the framework recommended by Van Gelder (1995) as an alternative to the computational one.[45] By taking computation out of the picture, most authors assume to be taking representations out of it as well. That's an unjustified move, though. The explanatory role attributed to representations in cognition may vary enormously. For instance, some take the role of causal mediation to be enough, while others require something that is able to work as a decoupable stand-in (i.e. a proxy) for something else (Ramsey, 2007). Whether these can be made to work is not at stake: the point is that you can't get rid of all of them at once just because you reject the computational framework.

Indeed, it is possible - at least in principle - to give explanatory roles for representations in dynamic models. I'm not aware of anyone who advanced such a project explicitly, but it is illuminating to recognize that this is contingent and the idea could be pursued. It is also remarkable that there is no consensus on the relationship between dynamic and computational systems. While van Gelder thinks that they are dissociated frameworks and completely different in nature (Van Gelder, 1995), others like Wheeler think that computational systems are a kind of dynamic system (Wheeler, 2005). This provides an important clue to better understand the dialectic of the debates between cognitivists and dynamicists. Many arguments for the adoption of dynamic models target only very specific and classical conceptions of computation and representation such as those that takes cognition to be the business of *physical symbol systems* (PSS) operating through heuristic information search strategies (Newell, 1976). However, there's a whole gradient of possibilities in between that remains untouched.

1.5.2.1 The Wheeler-Dreyfus debate on the role of representational contents

What's the place of Heideggerian cognitive science in such a gradient? That was the starting point of the debate between Dreyfus and Wheeler. Both of them accept that there can be some place for representations and computational mechanisms in the mind, that is,

[44] Dreyfus's reading is typically regarded as deeply flawed among scholars. But Wheeler's approach is based on his own reading, which motivated some debate with Dreyfus.

[45] For a nice review of this kind of models, see Chemero; Silberstein (2008) and Faries; Chemero (2018).

they're not against representations *per se*.[46] The disagreement is not about the whole edifice of cognition, but whether there can be any role for representations in explaining the relevance-sensitivity that underpins commonsense holism, situation holism and the capacity to negotiate non-saturable contexts. This is one of the reasons why the debate is primarily focused on the so-called online cognition. While Dreyfus thinks that online cognition must be representation-free, Wheeler's perspective is more-fine grained: he thinks that, in at least some cases, a particular kind of representation might be involved in online cognition, but he agrees that such representations have no explanatory role to play in our relevance sensitivity.

The boundaries of online and offline cognition are somewhat fuzzy, but given our current concerns, some stipulative remarks will do. The distinction revolves around different forms of tracking the world. Online cognitive abilities track it by relying heavily (or completely) on the coupling between the encompassing system and its surrounding environment. Such coupling is usually characterized in terms of the close interaction between perception and action. Thus, the primary focus of online cognition is that of controlling action. This kind of tracking may already involve some pretty complicated tricks, though. We're tracking our friend in a crowd, even though the visual inputs come only from a part of his body at any point in time, depending on how much of him is occluded by the crowd. If that friend becomes completely occluded while passing behind a truck, we can usually predict with reasonable accuracy that he's coming out on the other side, considering (for instance) its pace. In its turn, offline cognition is decoupled from the system's immediate surroundings, i.e. a fully (or mostly) detached negotiation with the world. It involves worrying about the weather somewhere else, thinking about someone who's not here or reevaluating a decision made long ago.

We can still think of it as a way to track the world, though. Rather than focus on the "here and now", offline cognition can account for a fully detached "there and then". If we lose our friend in the crowd, we can manage to find him again, for instance, by relying on the knowledge that he promised to be back home in two hours or maybe on what we know about its previous behaviour. That's why one should not think about online and offline cognition as a clear-cut distinction between two radically different kinds of cognition, but rather as the extremities of a continuum.[47] As we take things to be detached from the world, we open up new ways of handling them: what was previously taken as possibilities of action (somewhere to sit) can now be regarded as objects with properties and relationships (a chair).[48]

[46] If this claim comes as surprising for readers familiar with Dreyfus' work, just remember that Dreyfus' criticism against GOFAI and cognitivism aims at the idea that human intelligent behavior can be grounded in, or explained through, physical symbol systems. Higher cognitive abilities such as those presented in detached or offline cognitive capacities might well be explained or modeled in computational and representational medium, provided that such mechanisms operate inside a background of more basic cognitive capacities that accounts for our basic understanding of the world. He was wary of the idea that computation could explain the whole edifice of cognition, but would have no big quarrel with the idea that it is useful or accurate to explain certain capacities that emerge only on the 3rd floor.

[47] This is not a claim about the underlying wetware, i.e. I'm still not getting into the issue whether we need distinctive kinds of cognitive machinery or capacities in order to make sense of the distinction. The nature of the continuum is still an open question. The current point is just that there are many shades of gray in between the extremes, and we need to make room for them.

[48] Whether what we consider detached takes on the world need to be accounted for with resources that are

With this in mind, we can get back to the disagreement between Dreyfus and Wheeler. For Dreyfus, nothing truly Heideggerian (whatever that amounts to) can allow any role for representations in online cognition. Our basic understanding of the human world is all set up at this level by sets of abilities structured according to our needs. In a classical example: at this level, to know what a hammer is amounts to knowing how to use it, and this is connected to knowing how to use nails, as well as knowing how to use the fences one can build with hammers and nails, and so on. Our whole capacity to cope with things in the world is accounted under this tool-like logic. This is very different from carving out the world in terms of objects with properties and relations to other objects. To take things as substances completely detached from the practical grip that comprises online cognition is the business of offline cognition.[49] One can know a lot about hammers while regarding them as objects (shape, weight, color and so on) without having a clue about how to use them. For Dreyfus, offline cognition can be accounted for in computational and representational terms, but only provided that such processing takes place in the previously structured background of worldly knowledge brought about by online cognition. Such background already maps what is relevant and what is not in each specific circumstance. That's why, for Dreyfus, offline cognition must take relevance as a given from online cognition. RP is precisely what you get by trying to handle online cognition from the theoretic perspective characteristic of offline cognition.

Wheeler, however, won't go so far. He definitely agrees with Dreyfus that relevance-sensitivity is the business of online cognition. Yet, he thinks that there is a role for a specific kind of representation. Such a role, however, is different from that of classic cognitivism. As we all know, the core cognitivist idea is that sub-personal mechanisms represent states of affairs in the world and compute their dynamics. Thus, given a state of affairs s1, if one switches the place of any object, the new state of affairs s2 is represented after being computed from s1 by using knowledge about the relevant domain. In this case, such knowledge is propositional, traffics in symbolic medium and consists of explicit descriptive axioms and rules. That's classic cognitivism's way to account for the conception of the mind as a mirror of the world. The mind's thinking about the world covaries with it by computing it. In this picture, the role attributed to representation assumes that their contents are objective and detached, as if they describe the world from the viewpoint of nowhere.[50] A nice example of this classical approach can be found in the work of Marr (2010). Marr's focus was the visual apparatus. His challenge, as he himself conceived it, is that of explaining how the mind manages to produce objetive, non agent-o-centric representations such as "glass at location x,y", out of the relatively poor 2D contents comprising the inputs of the visual system. These were the kind of contents that Dreyfus rejected in online cognition. Both Dreyfus and Wheeler reject that this kind of representational content can be found in online cognition. However, while Dreyfus believed that this was a good-enough reason to reject any kind of representational contents

themselves decoupled from the system's environment (e.g. representations) is a problem for later.

[49] Dreyfus associates this to Heidegger's presence-at-hand mode of intelligibility.

[50] With a nod to Nagel (1986).

in online cognition, Wheeler thinks there's a role for other kinds of representations.

To get a first glance of what Wheeler has in mind, we can look again at Marr's work. Curiously, part of Marr's account involves an intermediate step in which the cognitive system produces "2.5D" representations of distal environmental elements from the vantage point of the agent ("the light pattern that characterizes the front side of the transparent glass at my right"). The perspective-free characterization of the visual input ("glass at location x,y") is produced in a further step. The apparent need to consider 2.5D representational contents only as an intermediate step was a product of conceiving the mind as an objective mirror of the world. However, proponents of 4EA cognition conceive the mind not as a mirroring device but as a tool to control the interaction between organism and world. That's Wheeler's conception, but instead of following Dreyfus and rejecting representations altogether, Wheeler follows Clark (1997a) and claims that there is a kind of representation that can play an explanatory role in the mind-as-control. I'll call them *situated representations*.[51] While a classical, perspective-free representation negotiates the world through contents such as "glass at position x,y", a situated representation has contents that only makes sense by assuming an encompassing system or environment, such as "glass is at a small hand movement towards there".

Wheeler and others have noticed that Dreyfus is blind to the distinction between situated and non situated representations.[52] His argument against representations is a variation of an argument we've met before when discussing heuristic approaches in frame systems: in order to find the relevant frame to apply in a given situation, a second-order frame is needed, but in order to know the relevant second-order frame, a third-order frame is needed, and so on.[53] Given the non-saturable nature of human contexts, there's no stopping this regress. But why would Dreyfus think that the issue can be generalized to every representational approach? The reasoning can be synthesized like this: with perspective-free contents, in order to represent that there's a glass on a table, we also need to represent the glass, the table, the room and whatever else is there, including the set of possible relations among them. Otherwise, representations cannot fill their role in explaining how the system can predict possible permutations of the current situation ("if the glass is pushed out of the table, it will fall on the floor and probably break"). Thus, every representational state or process seems to imply a fully detached take on the world. In its turn, this implies the need to fully represent the world in a perspective-free way, which brings us back to the problem of how to avoid getting lost in this sea of representations. In other words, the explanatory role that classic cognitivism attributed to representations is its only possible role.

The problem with this reasoning is that it ignores the possibility of situated contents. The claim that a representational state or process is situated implies, among other things, that its

[51] Wheeler and Clark calls them action-oriented representations. This is partially due to their focus on online cognition. However, as we'll see later, I think this kind of representation is also useful in offline cognition (which means they can be decoupled) and thus I wanted a name that could smoothly navigate between online and offline cognition models.

[52] See Salay (2009) for instance.

[53] Dreyfus presents this version in Dreyfus (2007), and Wheeler discusses it in Wheeler (2008).

content can be determined in a way that does not bring with it the need for additional representational resources. Thus, a tracking mechanism can use situated representations in order to cope with some aspect of the current circumstances without the need to represent the whole world encompassing it. Such a content will only make sense for a given system, with a given body integrated with its surroundings in a given way. But that does not undermine its explanatory role of how can the system exercise the respective cognitive capacity. In presenting this point, Wheeler uses a nice example of a watch we know to be non-accurate: we see that it marks X, but we know the actual time is, say, X+2. This is a way to track the world that may involve the appeal to representational states in order to comprise a kind of "augmented reality" based on which the system will perform. We don't need to posit a broadly cognitivist framework — like Fodor's — just to gain explanatory leverage through representational content.[54] Situated representations reminds us of that.

I understand that this claim might raise lots of questions about the semantics of representational states and its relation to the attitude contents we find in beliefs and desires. I'll have a lot more to say about that in the coming chapters. What's important for the present purposes is to see that Dreyfus's rejection of representational contents is grounded on a blind spot regarding the possibility of situated representations. This will enable us to isolate questions regarding the reliance on representations and focus on the second part of the debate: whether and how the adoption of dynamical systems can help us to avoid RP.

1.5.2.2 The Wheeler-Dreyfus-Rietveld debate on relevance sensitivity

The next topic in the Wheeler-Dreyfus debate is more directly related to the claim that a broadly Heideggerian cognitive framework would be able to avoid RP. Wheeler's bet is that we can achieve this by relying on two different kinds of couplings. We have already met the first one when discussing the role of context in cognitive explanations: special-purpose adaptive couplings (SPAC). As we have seen in the example of the cricket phonotaxis system, a SPAC is triggered only by very specific inputs which leads the organism towards a given behavioral output. Wheeler's point is that such couplings are tailor-made for a single context. In his own words:

> (...) rather than starting outside of context and having to find its way in using relevancy heuristics and so on, the cricket's special-purpose mechanism, in the very process of being activated by a specific environmental trigger, brings a context of activity along with it, implicitly realized in the very operating principles which define that mechanism's successful functioning. Here, context is not something that certain causal mechanisms must reconstruct, once they have been triggered. Rather, context is something that is always there at the point of triggering, in the adaptive fabric of the activated mechanism. (2008, p. 335)

[54] It is important to remark that this is not an argument *for* the adoption of representations. I'm not claiming that this kind of mechanism cannot be explained unless we posit representations. The claim is that representations *can* be posited in explaining it without triggering Dreyfus's worries. I do think that representational accounts are the most fruitful, but that's a matter for later.

Wheeler thinks that this kind of coupling is enough to explain our sensitivity to relevance within a lot of different contexts. The couplings comprising the human mind could be of course much more complex than the ones we find in crickets. Presumably, a human SPAC could allow sensitivity to information available in contexts of social interaction, for instance. This is compatible with non-representational approaches that rely on complex structures of ecological information or "information for action" (Carvalho; Rolla, 2020b). But that's also why Wheeler would see no problem in accommodating situated representational contents within SPACs whenever they show themselves to be useful. Indeed, whether a SPAC is best modeled computationally or using the Dynamic Systems framework is something that must be answered case by case. In this regard, Wheeler's characterization of SPACs is somewhat neutral. His picture is one in which the human mind is essentially (though not exhaustively) a large collection of couplings like these (Wheeler, 2005, p. 278), and that's why, at this point, one might already be wondering what is the difference between a Wheeler's SPAC and a Minsky's frame. The answer is a spoiler of where our discussion is going: as far as RP is concerned, not much.

The mind is not comprised of just SPACs, though. There is a second kind of coupling that is much more complex and also important in understanding how can humans be sensitive to relevance: *continuous reciprocal causation* (CRC). This kind of system is non-linear and complex enough to challenge the kind of functional analysis that's behind broadly mechanistic approaches. That's why the task of computationally modelling them is (allegedly) hopeless, and we need dynamical models. A nice metaphor to understand CRCs can be found in Clark (1997a): a CRC coupling is like a jazz trio. Each member is constantly responding to the other ones and influencing their responses. There is no blueprint or script to be followed, nor any of the musicians have a distinctive role in leading the others. Whatever leads the system in one direction or another is causally spread throughout the whole system in a non-trivial way. The core idea is that this kind of coupling is responsible for the continuous fine-tuning between brain, body and environment. Such fine-tuning would allow us to track local and distal environmental minutia and trigger the appropriate behavioral output in a relevance-sensitive way.[55]

But how exactly collections of SPACs and CRCs allow for the dissolution of RP? In Wheeler's analysis, RP is a two-headed beast and manifests itself in two distinct ways that Wheeler dubs intra-context and inter-context (Wheeler, 2008). The intra-context RP is manifest when determining the adequate behavioral outputs within an already identified context. The inter-context RP, however, is manifest in context-identification tasks. While SPACs guide cognitive processing within a context, a CRC is the kind of coupling that allows for the identification of the context in which the system is currently embedded in.

[55] In a variation of this idea, Vervaeke; Lillicrap; Richards (2012), claim that many CRCs can be understood in terms of opponent processing. The fine-tuning between the agent and its environment is a feature of the balance between competing goals chased by distinct CRC-like subsystems. This might be a nice move towards more accurate dynamic models. But the upcoming criticism is not addressed by Vervaeke's emphasis on opponent processing. See also Vervaeke; Ferraro (2013).

Before discussing the details of Wheeler's suggestions, it is instructive to appreciate how remarkably similar this picture is to the broadly Fodorian one we've previously discussed. We are again talking about a kind of sub-system that is somehow specialized in a given situation or set of situations (frames, modules, SPACs...) and we are again discussing how well these can fare with or without some kind of system that allows for context identification (higher-order frames, energy allocation patterns, central cognition, CRCs...). Right now, we can only wonder whether this similarity is substantial or superficial. But this is the core of the second main disagreement between Dreyfus and Wheeler. The question of whether there's a substantial distinction between an intra-context and an inter-context RP is identical to question of how SPACs and CRCs relate to each other.

Dreyfus considers the conception of RP as two-headed somewhat artificial: given that it is up to CRC couplings to explain how a cognitive system can cope with its circumstances in a relevance-sensitive way, there is no explanatory work left for no SPAC. Such mechanisms can account for other cognitive capacities, but only in a way that presupposes a solution to RP. Relevance is the exclusive business of CRCs. In his turn, Wheeler insists that SPACs and CRCs are equi-primordial, and that we need both of them to account for both horns of RP. On this point, the most interesting argument against Wheeler's position does not come from Dreyfus, but from Rietveld (2012). The argument is grounded on a pathology named utilization behavior (UB):

> It describes a phenomenon in which patients with a lesion of the frontal lobe (and/or of interconnected subcortical structures) demonstrate an exaggerated dependency on their environments in guiding their behaviors. Patients with UB grasp and use familiar objects when they see them, neglecting their current situational irrelevance (...). Such a UB-patient may, for example, put on a pair of glasses even though nothing is wrong with his eyes. Or upon seeing a bed he may start to undress, although this bed is in someone else's house. A light switch in his visual field may make him turn the light on and off continuously. (Rietveld, 2012, p. 111)

Rietveld's point is that mechanisms like SPACs cannot handle relevance in isolation. They can only afford UB-like outputs. In order to avoid that, SPACs must continuously articulate themselves with a established tuning between brain, body and world such as the one provided by CRCs. *Ergo*, CRCs have explanatory precedence when it comes to relevance-sensitivity. If the tuning is lost, as it seems to be the case with UB-patients, the triggering of a SPAC cannot be sensitive to the current contextual tonality anymore, resulting in inappropriate behavioral outputs. Thus, Wheeler's distinction between intra-context and inter-context RP is unhelpful, for if we can't isolate the work of SPACs from the work of CRCs, then we can't isolate intra-context issues from the broader inter-context ones. There is no such thing as a purely intra-context relevance-sensitivity, which means there's no such thing as a purely intra-context RP. In the words we've been using, Rietveld's point is essentially that, except in pathological cases, an m-context can never account for the whole contextual tonality in which the agent is embedded.

As I see it, Rietveld sheds light on a major issue with Wheeler's framework, but getting there will require some unpacking. For starters, we must not be misled by the issue of when should a SPAC be put to work, as Rietveld sometimes seems to be. Wheeler's favorite examples are probably to blame. When bringing up the female cricket's phonotaxis mechanism, for instance, Wheeler emphasizes the fact that it is driven by a very specific environmental stimulus (the male's mating song). This might mislead us into thinking that SPACs are only useful in cases where some external clue can impose itself over the agent's contextual tonality. In this view, like a Fodorian module, a SPAC would have mandatory processing: whenever a given input arises, the mechanism starts working and affects the agent's behavioral outputs accordingly, regardless of the system's current contextual tonality. This, however, would amount to an account of context-switching, for we'd be claiming that, given a certain stimulus, the system will necessarily switch to a certain m-context. This is incompatible with the claims from both Wheeler and Rietveld. They explicitly take context-switching to be the business of CRCs, which means SPAC triggering cannot be the issue.

At this point, one can argue against Wheeler and Rietveld that there is evidence of specialized mechanisms "taking over" context-switching tasks. For instance, there are cases in which actions can be primed directly by objects in a way that mitigates (or even precludes) the sensitivity to the current contextual tonality.[56] In particular, Riddoch; Humphreys; Edwards (2000) show that the process of selecting a visually present object out of others is functionally distinct from selecting a response (by choosing the hand with which to grab a tea cup, for instance). Once both the object and the action get selected, normal-functioning agents are usually able to preempt the opportunities for action provided by other visually present objects, while in UB-patients this capacity degrades. These findings suggest a role for SPAC-like mechanisms in visual object selection that precludes an antecedent role for the relevance-sensitivity (supposedly) provided by CRCs. In the resulting picture, a failure to preempt irrelevant visual affordances for other objects would amount to an m-context that is imposing itself as a direct consequence of some given stimulus.

What do these findings mean for the claims of Rietveld and Wheeler? As far as the role they attribute to CRCs in context-switching goes, I don't think they're much of an issue. The mere existence of some mechanisms resistant to the priming of the current contextual tonality is not a threat to none of their pictures. Quite the opposite, it's something to be expected, given the evolutionary history we share with other species and the prominence of this kind of mechanism in them. New findings might change the view on how much of our normal behavior is a product of this kind of mechanism, but we already know that, at some point, an increasing number of mechanisms must start articulating with each other in increasingly complex ways. Otherwise, the task of accounting for commonsense and situation holism is hopeless. Remember, for instance, that one of the consequences of non-integration in frameworks conceiving the mind as a massive collection of specialized sub-systems is the need to massively redupli-

[56] See Riddoch; Humphreys; Edwards (2000) and Phillips; Ward (2002).

cate the information in each of them.[57] CRCs are an answer to this need. The moral is that, at least in a vague evolutionary sense, the evidence gives a small advantage to Wheeler's picture of SPACs and CRCs as explanatorily equi-primordial. We know specialized mechanisms to be present in other species, and it's more evolutionarily parsimonious to assume that, when it comes to us, nature kept improving on them (e.g. by allowing its increasingly flexible and context-sensitive application) without getting rid of its core strategy.[58]

The upshot is that the SPAC's reliance on CRC for triggering presents no real issue for the distinction between intra-context and inter-context relevance-sensitivity. In order to play its role adequately, a SPAC must have been previously fed by inputs from whatever prior processing took place. In other words, the holistic organization of the system always have a role in setting the boundaries of the encapsulated domains in which SPACs will take over. Even further, the SPAC itself might be a product that emerges on-the-fly out of the holistic transient phases in a system. Wheeler's general point is that once such a mechanism gets triggered, it can exhibit relevance-sensitivity within a given causal domain (what we have been calling a m-context) in a way that is insulated from other mechanisms. Once s SPAC is established and triggered, it can bear the explanatory burden of relevance-sensitivity within its underlying causal domain. This is the explanatory role Wheeler attributes to m-context, and this is why Wheeler argues that there can be a relevance issue for insulated SPACs, after all. By claiming that UB-patients are affected on whatever underpins the interaction between SPACs and CRCs (Wheeler, 2012), one need not commit to an explanatory precedence of some kind of mechanism over another.

Nonetheless, Wheeler's account is as far from hopeless as it is from safety. In order to fully avoid Rietveld's hit, he must show that SPACs buy us something that the sole appeal to CRCs cannot. We must be able to get some explanatory leverage by appealing to the notion of m-context as an insulated causal domain underpinning a special-purpose sub-system. Wheeler's claim is that if we recognize the possibility of handling relevance within a specific m-context by means of one or more SPACs then, at least in that m-context, RP will not be an issue. If we give up on the distinction between intra-context and inter-context relevance sensitivity, this possibility is lost and the whole project of a Heideggerian cognitive science of relevance-sensitivity becomes nothing but a non-computational guise of the same old RP. But can SPACs really buy us this? To avoid Rietveld's challenge, Wheeler needs to show how a cognitive system can establish the boundaries of what will become an m-context.

This is a well-known problem. Take the example in which the UB-patient can't help but start to undress and get ready to sleep whenever she sees a bed, even if the bed is at a friend's house. A reasonable assumption is that the UB-patient is stuck in a kind of *getting-ready-to-sleep* m-context. But how are we to tell whether a given feature should be considered part of it? Should the m-context include sensitivity just to the bed status (is it made?), or should the patient's current clothing get considered as well, so that it matters whether the patient is

[57] As we learned from the discussion of the massive modularity hypothesis.

[58] See Griffa *et al.* (2022) for an account of the evolution of information transmission in mammalian brains.

already wearing pijamas or not? Should it be sensitive to the location of the bed (at home, at a friend's, at a house party...)? What about the bed size, the temperature in the room or the surrounding noise? The familiar difficulty that underpins such questions is: how far should we go into framing this m-context? In order to account for the non-saturable nature of human contexts, we might end up encompassing pretty much everything the system is sensitive to and devoiding m-context of the role it's suppose to play. This kind of difficulty is the same one that a frame system faces, and nothing about SPACs suggest they can fare any better. In a nutshell, Wheeler needs nothing short of a theory of context determination. He is well aware of this. In his words:

> This is, I admit, a genuine risk, and one that casts a spotlight on the Herculean theoretical challenge of giving a theory of how our cognitive systems determine where the boundaries of contexts lie. It seems to me, however, that at this juncture we confront a choice. Either we remain committed to the distinction between intra-context and inter-context sensitivity to relevance, and so humbly accept the challenge to give a theory of context-determination, or we follow Dreyfus and Rietveld in rejecting that distinction, and so confront the fact that, (...), our Heideggerian cognitive-science of relevance-sensitivity in general, and not just of context-switching, has barely begun. (Wheeler, 2012, p. 209)

Wheeler's fear seems to be that if we accept a framework in which CRCs carry all the explanatory burden, both for intra-context and inter-context relevance sensitivity, then we'll gain no insight on how a cognitive system can be sensitive to relevance. We can see this worry in his reply to a critique of his framework staged in Dreyfus (2007). By relying on the work of Freeman (2001), Dreyfus presents his view on how CRC systems could account for relevance-sensitivity in cognitive systems: the whole horizon of possible behavioral outputs is rebuilt at each attention shift. In Dreyfus's words:

> (...) our sense of other potentially relevant familiar situations on the horizon of the current situation [i.e., our readiness and capacity for context-switching], might well be correlated with the fact that brain activity is not simply in one attractor basin at a time but is influenced by other attractor basins in the same landscape, as well as by other attractor landscapes which under what have previously been experienced as relevant conditions are ready to draw current brain activity into themselves. According to Freeman, what makes us open to the horizontal influence of other attractors is that the whole system of attractor landscapes collapses and is rebuilt with each new rabbit sniff [Freeman has worked extensively on rabbit olfaction], or in our case, presumably with each shift in our attention. And after each collapse, a new landscape may be formed on the basis of new significant stimuli – a landscape in which, thanks to past experiences, a different attractor is active. (Dreyfus, 2007)

Dreyfus develops his point using words from Dynamic Systems Theory, but the essential claim is that at any given time we're not only operating under the influence of whatever comprises the current context, but also under the influence of whatever was previously experienced in similar circumstances. According to Dreyfus, this is possible because, unlike frame systems (or representational systems in general), we need not rely on previously built links (or "routes") between frames. In analogy, the resulting picture is one in which, at each attention shift, the whole frame system is rebuilt by taking into account everything we're familiar with. That's why a *restaurant* situation can easily become a *restaurant-with-co-workers-around* as soon as the agent spots any co-workers.

In my view, the problem with Dreyfus's suggestion is that it begs the question of relevance-sensitivity. His move seems plausible for those not very fond of computational systems because to rebuild a whole frame system from the scratch seems computationally unrealistic (i.e. there is no algorithm for it). So it might seem that dynamical models are buying us something in this case. But they're not.

Computational systems can easily reorganize large amounts of information according to certain principles. Every time you search something on Google, it's engine is reorganizing the huge amount of information it has in a way that is connected to your query. The million dollar question is: what is the principle (or the tractable set of principles) that accounts for such reorganizing in a way that is sensitive to what's relevant at any given circumstance? Frame systems, as we have seen, are a kind of answer for this question. Unfortunately, in the absence of the aforementioned principle, we cannot create a frame system that reorganizes itself constantly by the lights of the most recent input or the most recent attention shift. All we can do is either 1) try to design one specific frame system and try to cope with worldly variations by using heuristics (this is the strategy we discussed in the previous session, and we already know why it is hopeless), or 2) design one whole frame system for every possible circumstance, which is not only unrealistic, but also unhelpful, for now we would need a meta frame system in order to guide the selection of the frame system that accounts for the current circumstances, and for that, we would need either that very same set of principles that would account for relevance sensitivity, or a second-order meta frame system. This is nothing but a new version of the regress we have met when discussing frame systems in the previous session. Google can do this by artificially constraining the possible relations between information it has on us and the information it makes available for search. If your query implies an unavailable articulation ("from all the books you know, give me the names of those containing at least three typos in its introduction"), it is very likely to fail, even if the information is there.

Wheeler seems to share enough of my view to pressure Dreyfus on precisely this point: his suggestion begs the question because he lacks an account of how the rebuilding of the whole landscape of possible behavioral outputs can be sensitive to what's relevant. If the system is, at any time, sensitive to everything that has been significant in the past (that is, everything the system is familiar with), how is it that the system is able to select only the relevant portions of it? Wheeler's pressure is akin to mine: simply pointing out that this view results in a flexible system is not enough, for we need to specify some kind of strategy or principle the system applies in order to cope with relevance. In its absence, all we can do is reformulate the problem in a new framework, and that's what Dreyfus did.

To claim that CRCs enable us to avoid RP is not really that different from claiming that heuristics are all we need in computational frameworks. As we have seen, heuristics represent the hope that RP is all about avoiding computationally unfeasible processes. But that falls short of what's necessary, for heuristics must themselves rely on some principled and relevance-sensitive way to organize the information it works with, and we don't know how to do that. In the same vein, CRCs represent the hope that RP is all about mapping complex mind-body-world relationships. Though these are necessary, they're far from sufficient. We could formulate dozens of heuristics and still have no clue as to how the system knows which one to apply in an open-ended set of non-saturable contexts. The same goes for CRCs: we could describe dozens of it, and still have no idea how can their interaction be sensitive to relevance. Thus, in order to avoid Dreyfus's fate, Wheeler holds tight to its distinction between intra-context and inter-context RP: given how clueless we are regarding CRCs, SPACs are all we have left, but they're only helpful under the assumption that RP has an intra-context face.

However, it is not clear whether Wheeler's credit card has a big enough limit. Even if the issue of m-context determination were handled properly, we'd still be lacking a substantial improvement over the reliance on CRCs. SPACs rely on bounded causal domains, which amounts to determinate m-contexts, leaving the non-saturable nature of human contexts out of the picture. This is not a problem *per se*, for it would be implausible to assume a 1-1 mapping between every SPAC's m-context and each agent's contextual tonality. There's probably no such thing as a restaurant-SPAC handling relevance within restaurants. A restaurant context is likely to be the product of many narrower m-contexts handled by different SPACs. But in order to account for full-fledged human contexts, we do need a way to explain how can many m-contexts add up and constitute the agent's contextual tonality.

The problem is that we still don't know how to build systems that are sensitive to non-saturable contexts, and it's a wide open question whether this can be achieved through the complex articulation of sets of m-contexts, be they data structures like frames, or bounded causal domains underlying SPACs. This is, again, the issue of how to organize cognitive resources. RP, remember, is not just about finding *some* way to organize resources. It's about explaining how a system can find the *right* way to organize them, and by that I mean the one that allows the system to apply those resources in a way that is sensitive to relevance. Furthermore, the right organization must not only account for contexts the agent is familiar with, but also for her context productivity. In order to cope with new situations, the agent must be able to extrapolate from the contexts she already knows in a way that is relevance-preserving. In a nutshell, this is the very same challenge that Dreyfus and Rietveld face by relying solely on CRCs. Consequently, as far as RP goes, SPACs bring no new game to town, which means they fare no better than CRCs and Wheeler's distinction is not really buying us anything.

1.5.2.3 Taking non-representational stock

It seems like, at the end of day, friends of dynamic models cannot consider themselves free of RP, for it is neither representational nor computational in nature. Even if dynamicists are able to find a set of principles or strategies capable of accounting for how we apply cognitive resources in a productive yet relevance-sensitive way, I see no reason why there could not be a computational guise of such a set. It can go the other way as well: we can probably produce a dynamicist guise of any computational solution that eventually arises. This might be somewhat obvious for those who think like Wheeler (2005) about the nature of dynamic systems. For him, computational systems are a kind of dynamic system. But even if one follows Van Gelder (1995) and conceives dynamic and computational systems as different in nature, my point remains. This is not exactly a problem for most dynamicists, for it might be seen as just one more item in a long list of ongoing debates with representationalists. It is, however, a hard hit on Dreyfus, for RP is his core motivation for rejecting representations and computations in online cognition.

Nonetheless, one can of course insist on the bet that relevance sensitivity will turn out to be the kind of capacity that resists psychologically realistic computational solutions. Take CRC systems: as we have seen, they model a nontrivial causal spread between brain, body and environment that usually resists functional analysis and mechanistic descriptions. If relevance-sensitivity turns out to be produced by this kind of system, then even if we can model it computationally, the result won't be realistic, in the sense that it won't throw any light on how creatures like us manage to cope with relevance. Though this is possible, I don't see it as an explanatory virtue of the Dynamic Systems framework over the computational one. Given any dynamic model, even van Gelder's now famous Watt Governor (Van Gelder, 1995), those more mechanistically inclined (like me), can always say "nice, but how does it work?", that is, we can still ask for an explanation of how such a capacity comes out of such a complex causal interaction.

Dynamic models, as I see them, cannot answer that. They allow for a kind of good science by subsuming causal interaction data under some regularity or law, which is enough to formulate counterfactual hypothesis. However, as Cummins (2010a) remarks, this is an exercise of confirmation, not explanation. The data confirms the law or, equivalently, the law describes the data, but it doesn't actually explain it. Remember the aforementioned McGurk effect: to describe the conditions under which it occurs is not to explain it, but only to describe it. This is different from explaining capacities in terms of the interaction of simpler capacities, that is, by accounting for the behavior of some system in terms of its constitution and organization. Such descriptions are psychology's *explananda.*[59]

That's why if we eventually reach a place where there is a given psychological capacity (such as relevance sensitivity) that cannot be explained in terms of the constitution and organization of its encompassing system, I'll take this to mean that it cannot be explained at all (given the currently available set of explanatory frameworks, of course), and not that it can only be explained through a Dynamic Systems framework. This is why friends of dynamic

[59] Claiming that subsumption under law cannot be explanatory in psychology is undoubtedly disputable. I have no intention to dwell on that dispute, however. My intention is just to be clear on what grounds I hold my position about this matter, so the disagreeing reader and I can both know what we're not agreeing about. See also Cummins (1983) and Haugeland (1998b).

models are not allowed to say that, in their framework, RP simply does not emerge. RP can only emerge within an enterprise that is trying to explain how something works. If that's not possible, then just like the fundamental laws of physics, we'll be limited to only describe how things are. However, I think there is reason for some optimist in providing a psychological explanation of relevance sensitivity, and that is what motivated this work.

1.5.3 What about this nice framework here?

I can hear a lot of questions already. What about Noë's actionism (Noë, 2004, 2012), *autopoietic* enactivism (Varela; Rosch; Thompson, 1991), radical enactive cognition (Hutto; Myin, 2013), the ecological approach from Chemero (2009) or Bayesian approaches such as those endorsing predictive processing (PP) (Clark, 2016; Hohwy, 2014)? The analysis provided in the previous sections were (I hope) instructive, yet far from exhaustive. *Pace* Fodor, there are just too many games in town to allow for a comprehensive analysis of each one of them. This is why I've tried to show how the same problem can emerge in two extreme cases: on the one hand, the most classical version of cognitivism, and on the other hand, a Heideggerian approach that purports to be radically different in rejecting both computational mechanisms and any explanatory role for representations, at least as far as relevance sensitivity goes. If both frameworks cannot cope with relevance, then they can't cope with the non-saturable nature of human contexts, which in its turn shows that our capacity for commonsense holism and situation holism remains beyond their explanatory reach.

Can we generalize this conclusion to all available frameworks? I think we can, but let me elaborate on this. I'm not claiming that nothing about any of then can be helpful in eventually finding an alternative solution to RP. I'm also not claiming that the best solution eventually available will not rely on anything specific to these frameworks. It might be that the best approach involves something that is available to only one of them. Indeed, possessing an exclusive conceptual tool that enables a solution for RP would be, by my lights, a huge motivation for its adoption. What I am claiming, however, is that no framework can consider itself free of the relevance issue just in virtue of its core tenets. All of them owes an explanation of how could RP be solved in its own terms and, assuming that's the case, arguments for accepting that the solution is not available to other approaches. Right now I think it's very unlikely that any of them has such a conceptual tool or a proprietary solution. In fact, I think that's also true of the suggestion's I'm going to develop in the next chapters. All of them can be easily adapted to (for instance) PP frameworks like the one advocated by Clark and Hohwy. One could even wonder why I didn't do it myself. The answer is simple: because I think that RP should not be regarded as a framework-specific issue. It is true that, to some extent, the claims I'm gonna make will be incompatible with some frameworks (Fodor's cognitivism, for instance) and will thus imply their rejection. But I still think it's something good to be as framework-neutral as one possibly can.

Having said that, it might be useful to make a few quick remarks on why some well-known approaches or ideas could mislead us into thinking that they'd have an easier time coping with RP. The history of RP teaches us that there are many ways to underestimate it. Thus, whoever is serious about trying to get somewhere new must build on this story. Furthermore, it is always fruitful to do our best in mitigating the amount of readers going through all the work wondering "why in the world wouldn't X be enough?".

Let's start by considering the fact that situated cognition allows us to offload much of both cognitive processing and information to the environment. We have seen that the issue is broadly organizational. Well, isn't embedding a way to organize? To play pretend in a stage is a way to organize what's in and what's out of the play. To put some toy cars in line is a way to organize them in a street-like structure that will enable street-wise inferences. To be in a classroom is a way to prime the need to abide by certain rules. Couldn't we just let such structures in the world leads us towards what's relevant and what's not? I'm sympathetic to that idea and that's exactly what I think happens in some cases. But I don't think it constitutes an overall solution to RP. Much of our cognitive tracking relies on the kind of informational integration that characterizes our commonsense or situation holism, and it's up to us to articulate whatever is out there in a relevance sensitive way. Though sometimes external informational structures may impose themselves on us, in most cases, our cognitive apparatus allows us to articulate them with our goals and expectations in order to predict outcomes *from* them, as well as to render and exploit permutations *of* them.

Even further, we can build on complex, higher order, articulations of such predictions and permutations. The need to be sensitive to relevance encompasses these abilities as well. While being in a classroom may help to constrain inferences that are foreign to that kind of context, we don't lose the ability to overcome the constraints nor the ability to enforce classroom-like constraints in other places. As we have seen, utilization-behavior patients illustrate what happens whenever this capacity of us gets degraded or lost. In a nutshell, we can't be regarded as puppets of external informational structures, no matter how rich or complex they are. Our relevance sensitivity is manifest in what we make of them, for what we make of them frequently involves commonsense and situation holism. Taking the system-environment as the primary unity of analysis throws no light on how we do that.

Thus, as far as RP goes, it doesn't really matter whether those informational structures are somehow encoded within our heads or available out there. RP is about exploiting whatever resources we've got in a relevance-sensitive way. Additionally, the difficulty in explaining this capacity to exploit is not really about the amount of processing one has to do, so offloading it won't buy us much. If the problem was about processing effort, non-computational approaches could fare better, and we already know that, and why, they don't. Finally, one should not forget that the worry about RP encompasses both online and offline cognitive capacities. We're not just trying to explain how creatures like us can cope with changes in the world, but also how we can reason about a changing world. Whenever the world is absent, we'd better have something in its place. And that something, whatever it amounts to (representations, know-how, etc.) constitutes the set of cognitive resources that must be managed in a relevance-sensitive way.

The second remark I want to make is about affection-guided cognition. I have no doubt that cognition is affective in the very general sense that we care about (at least some of) our inputs and some of our processing, and that this changes the inferential paths we tread. Indeed, Haugeland made use of this idea to criticize AI efforts in modeling capacities that rely on commonsense holism and situation holism: *"The trouble with computers is that they don't give a damn"* (Haugeland, 1998a). We surely could make Deep Blue behave like it cares whether it wins or loses when playing chess with Kasparov, but this would be nothing but a scam. Even if we accept the claim that an artificial gamer has a goal, there's no reason to accept that it cares whether that goal is achieved. Since machines don't care, and it doesn't seem likely that we'll be able to make them care in a foreseeable future, I agree with Haugeland that this is a key limitation for AI. But I don't think affections are enough to handle RP in human cognition, even if we recognize it as broadly affective. As Carvalho (2019) and Ransom (2016) have shown, emotions are largely context-sensitive. Contextual factors might make one regard a given object as useful or threatening. Thus, affective marks, as far as cognition goes, must rely on previously identified contexts. Even if they eventually modulate context identification tasks (as when somebody's melancholic mood leads to non-usual processing and different behavioral outputs), this leaves no place for them in answering the key question about relevance sensitivity.

The final remark is about learning. It's a very special kind of cognitive capacity, and it will be given an important role in the upcoming chapters (specially in the last one). Learning has a special appeal when discussing relevance issues, for whenever we think about commonsense and situation holism as fully developed capacities, we usually figure the problem as one of framing or selecting a subset of relevant resources out of a huge pool of available ones. When thinking about learning, however, such top-down approach is usually replaced by a bottom-up one, in which we think about the developmental stages through which those resources get established. This leads to a legitimate question: couldn't we learn about what's relevant as we learn how to cope with the world? Couldn't relevance be out there to be learned, just like everything else?

Now, there is a trivial sense in which this is evidently true. We do learn what's relevant as we learn our way in the world. Unfortunately, in order to account for how we do that, we need a solution to RP. Just like any other cognitive capacities, learning needs to be tamed. It requires the right set of biases and constraints on the processing. We know this from the so called "no free lunch" (NFL) theorems (Wolpert; Macready, 1997). They established that there can be no general purpose learning strategy: every approach is necessarily tuned to a subset of problems. The more biased a learning strategy is, the more it assumes about its target domain, and the less it allows for variation. Conversely, the more variation is allowed, the more likely is the admission of noise in the learning trajectory, which means that irrelevant features are being taken into account. Thus, in order to learn how to cope with something, the system needs to apply the right strategy, with the right set of biases and the right amount of tolerance for variation. But how can we know the right learning strategy for a domain before we know anything about that domain?

Needless to say, the problem gets specially acute in trying to account for creatures like us. Our ability to handle non-saturable contexts involves the capacity to learn about an unbounded set of domains. But we can't have an unbounded number of domain-specific learning mechanisms, and the NFL theorem seems to rule general learning out of the picture. Thus, we're stuck with the familiar need to pick the contextually right learning strategy, i.e. we need to modulate our learning parameters in a way that render the process sensitive to what's relevant in the current circumstances. Notice how the problem with learning mechanisms resemble the issues we've found when discussing frame systems and Heideggerian cognitive science. We're again trying to make cognition sensitive to what's relevant with the same see-saw between untamed general processing or a huge list of very specific mechanisms that are suitable only to handle saturable contexts. I won't dig the issue further right now, for it will be properly discussed in chapter 4, where I'll develop my suggestion of how we can learn how to learn about an open-ended set of domains, i.e. learn to learn what's relevant. What matters for now is to realize that, just like the other alternatives we've considered, learning cannot account for RP on its own.

1.6 Mind the gap

RP stands as an obstacle to explanations of our capacity to handle non-saturable contexts. Frame systems, special purpose couplings, physiological markers, and other varieties of these approaches were all considered and dismissed due to their incapacity to account for non-saturability. With this in mind, after reading a previous version of this work, Marco Aurélio Alves and Eros Carvalho raised the following question: what if non-saturability is too strong a requirement for human-like commonsense?[60] Perhaps the difficulties we're facing are actually given to how the phenomenon is being conceived. Here's an example of a line of reasoning that could lead us into thinking that: we surely can write down an indefinitely large number of contexts with an open-ended set of elements. These elements may come from virtually any domain, which results in the isotropy emphasized by Fodor. But that's a feature of natural language. Why should that be taken as a clue for what's really going on within our cognitive processes? There's the possibility that we never take into consideration the hundreds of articulations comprising a given linguistic description of a human context. Human contexts may be relatively more complex than those involved in the cognition of non-human animals, but perhaps they're equally saturable and the big difference lies in the availability of more memory and processing power. Therefore, though it might seem like we can integrate cognitive knowledge from any domain to any other domain, this is an illusion. Hard limits to what domain and how exactly they can be articulated may apply.

[60] I'm very grateful to both of them for much helpful discussion of the ideas presented here.

I think what makes this reasoning tempting is actually a conflation of two distinctive cognitive capacities. First, the capacity to integrate information from many distinctive cognitive mechanisms that handle different domains. Second, the capacity to take whatever results from such integration and extrapolate from it without loosing track of what's circumstantially relevant. Artificially saturated situations (or saturated contexts from sufficiently simple creatures) seems to go without the second capacity precisely because the amount of information involved is small enough for us (or the creature) to handle it by exhausting all the available possibilities. But the need for the second capacity does not emerge only when the system is able to integrate everything to everything. Evidently, the more domains are available for integration, the more complex articulations we can produce, and the more RP bites when trying to extrapolate from it. And humans can render much more articulations than it would be feasible to exhaust, given our seriously limited resources. Yet we somehow manage to zero in on the relevant ones, even when it involves an open-ended set of distinctive permutations of contexts, possibly involving parameters that we never faced before. But even if we were capable of cognizing only a few small domains and render poor articulations of them, this would bring us no clue as to how we can apply that knowledge in new situations without losing track of what's relevant. Non-saturability is to be regarded as an extra degree of flexibility that creatures like us come to have.

In this view, rather than an "all-or-nothing" integrative capacity, we could imagine something more gradual. The image of "islands of rationality" endorsed by Hurley (2006) is useful here: perhaps we have increasingly complex island-like clusters of integrative capacities that provide sparse degrees of generality and flexibility among an increasingly larger number of and situations. Indeed, that's exactly how I see it. The point was never that our cognitive machinery can integrate any cognitive knowledge with any other cognitive knowledge, even though some authors (like Fodor) wrongly take that to be the case. RP is able to bother us much before that. Big islands of integrative capacities are enough to raise it, for they already challenge our time and memory constraints and thus preclude an exhaustive consideration of every permutation. The capacity to integrate information from different domains and exploit the resulting articulation with a variety of tools are not due to a single cognitive achievement. But there's a lot to unpack before I can tell the whole story. In chapter 2, we'll see how distinctive capacities can be integrated through different approaches such as encoding, problem-embedding, representational redescription, etc. There's nothing exclusively human about these approaches, though it might be true that we employ them at a higher rate and with a larger volume of information. In chapter 3, we'll discuss how this kind of strategy can be applied in order to render a gradual version of the gap between human and non-human animal cognition without the need to posit any mechanism peculiarly human Finally, in chapter 4 we'll see how we can make use of these tools in a way that (hopefully) won't beg RP (spoiler: it relies heavily on cultural evolution). If the story as I'm trying to tell it is in the right track, the gap between human and non-human animals is much more fuzzy than we're used to thinking of it. That doesn't mean that there are no important differences, but rather that what makes us human is perhaps much more of a fragile setup than we would readily admit.

1.7 Representational and inferential productivity

What do we got so far? RP is a hard, framework-neutral, non-computational and non-representational issue. It haunts whoever is trying to understand cognition within a naturalistic perspective. Ubiquitous issues of this kind require that we either account for it or live with it. Despite the difficulties it ensues, RP is relatively easy to live with, for it only bites those worried with the big picture. Most of the current research efforts among cognitive scientists are relatively narrow, though. They're usually focused on the details of a single cognitive capacity or maybe small clusters of abilities. Therefore, the problem of how can these be integrated in a coherent whole is rarely (if ever) faced.[61] Even if a solution is still out of reach, understanding RP is already rewarding, for it improves the knowledge of how close we are to provide a scientific explanation of the mind. To know what's being left out can be informative and useful. This is true not just of cognitive science, but any enterprise relying on similar conceptual and modeling tools as well. For instance, an understanding of RP throws light on what underpins crucial deficiencies of large language models such as OpenAI's GPT family: at least so far, no such model has a reliable grip on what's relevant in human contexts.[62] The same understanding justifies a lot of anxiety, for we know a bit more about the challenge lying ahead. Any explanation of human cognition must account for the sensitivity to what's relevant in an open-ended set of non-saturable contexts. If we can't do that, we can't explain commonsense nor situation holism.

It's time to check whether and how something can be done about it. I will outline now the strategy and the background employed in presenting the ideas comprising the upcoming chapters. I intend to stay as framework-neutral as possible, but I'll adopt a representational approach. Not any conception of representation, though. In chapters 2 and 3 I'll argue that symbols and linguistically inspired approaches to representation are a dead end. But that's not true of structural representations (Swoyer, 1991). In the market of ideas regarding content, structural representations buy us a lot of explanatory leverage for a very low price. Its purchases will (I hope) become increasingly clear as we advance. I can already remark, though, that they provide a clean and non-problematic link between online and offline cognition even for those - like me - willing to embrace 4EA approaches, i.e. the mind is embodied, embedded, (perhaps) extended, enactive and affective. We can have it all without giving up on the additional explanatory dimension that representation buys us.

This extra dimension comes from the fact that representations provide a clear-cut distinction between what we know about the world and how we exploit this knowledge. As we'll

[61] Not to be confused with worries about the disunity of scientific approaches. The disregard for the big picture that I have in mind could remain untouched even if cognitive science were to rest on a single framework.

[62] I wrote about it in Barth (2021), though in another connection.

see, this enables the formulation of empirical hypotheses that would not be available otherwise. When trying to get somewhere, there's a difference between making a mistake due to an inaccurate map, and due to exploiting an accurate map upside-down. Representations enable the same kind of distinction within sub-personal mechanisms. But this leverage comes at a risk, for an additional hard problem lurks representational systems: how do we store what we know in a way that allows for efficient yet flexible retrieval in an open-ended set of non-saturable contexts? In the next chapter I'll argue that, despite appearances, this is not simply a new instance of RP that emerges only in representational systems. It is a distinct issue called *frame problem* that needs to be handled through a different approach. This gives us two distinct *desiderata* that, for the sake of clarity, I'll call *representational productivity* and *inferential productivity*.

Representational productivity is about the expressive power of the representational architecture (or the many, as we'll see) employed by the mind. The set of things the human mind can think about is unbounded and that seems to require a representational architecture with (potentially) unbounded expressive power. This is, in part, what's behind Fodor's language of though hypothesis (LOT) through which we can account for a huge expressive power even with a small lexicon (Fodor, 1980). LOT seemed necessary for the cognitivist picture because it allows one to express how things are and how things will be after some event or action. It appeals to compositionality in order to account for how bounded creatures like us can exhibit unbounded expressive power. A problem with this approach is that LOT can express not just the proper outcomes of some event or action, but also every outcome that's logically conceivable, no matter how alien to the current circumstances (e.g. balls falling up) or even to the agent's world (e.g. balls becoming invisible). The only available choice for LOTers is to restrict the possible inferences a system can do with its expressive power. Modules and frames are a way to do just that. As we've seen, however, the cost is huge: if we try to tame what the system can express by constraining the ways it can exploit its representational resources, we pay the price in flexibility, which is essential to handle non-saturable contexts. In other words, LOT conflates representational and inferential productivity, i.e. it conflates the system's capacity to represent with the system's capacity to exploit what it represents. The distinction between using a wrong map and using a map wrong cannot be made to work in LOT. This point will be further developed in the coming chapter.

The alternative that I'll suggest is called *representational cognitive pluralism* (RCP). The core idea is that much of our limitations are due to the presence of different domain-targeted representational architectures (or, as I'll call them later, representational schemes) with varied representational power. Thus, our expressive power does not come from a single maximally expressive architecture, but from the complex articulation of many less powerful ones. In the resulting picture, we get a representational productivity that leaves no space for the representational aspects of RP. The idea will be presented throughout two chapters: the first focuses on conceptual issues, and the other is about its psychological and scientific applicability. If everything goes as planned, this will leave us free of any worries about representations in cognition (as far as relevance sensitivity goes) and in a way that is not restricted to online cognition.[63]

What about inferential productivity? That's just what we've been discussing throughout the chapter. It is the capacity to exploit the cognitive resources we have in a relevance sensitive way in an open-ended set of non-saturable contexts. By "inferential" I don't mean anything logic-like nor anything essentially related to language. I mean simply the processing comprising the exploitation of some cognitive resource (representations, detectors, know-how, etc.). My suggestion on how we should cope with this problem will be presented in the thesis's last chapter, after introducing all the necessary conceptual tools.

1.8 Paragraph-by-paragraph summary

Each entry below summarizes a paragraph of the main text.

What intelligence amounts to

Definitions of intelligence were always regarded as a secondary matter.

No wonder the term is being used in so many senses.

These differences might be radical, but not necessarily.

I'll take intelligence to be a cluster of capacities that allows one to deal reliably with more than the present and manifest.

Intelligence in this sense is manifest both in online and offline cognition.

I'll focus on two of intelligence's constitutive capacities: commonsense and situation holism.

Commonsense holism

Commonsense holism is the capacity to use our background knowledge about the world.

It is manifest even in the comprehension of simple stories.

It is holistic, for most of our knowledge is potentially relevant to handle an ongoing situation.

Though linguistic examples are the most common ones, we can see commonsense operating in perceptual tasks as well.

Importantly, being holistic is not about actually bringing to the fore as much knowledge as one can at every single task: most of the time, most of our knowledge must be intelligently ignored.

[63] I have no intention to claim that non-representational approaches are hopelessly trapped in online cognition. My point is just that this is a wide open issue that will not emerge for representational approaches such as the one suggested here. There are many possible ways for non-representational frameworks to account for offline cognition. Some like Kiverstein; Rietveld (2018) try to reconceive them in order to make it easier to handle with the available non-representational tools. In such a scale-down approach, the boundary between online and offline cognition is somewhat blurred. Others accept the traditional online-offline distinction and try to show how whatever tools their framework got can be used in both kinds of cognition. These can be regarded as scale-up approaches. For a nice review, see Carvalho; Rolla (2020a).

We can do that because we can take the world to be in many distinct ways at once, and we apply our cognitive knowledge accordingly.

Furthermore, it happens as the inputs come, for commonsense holism is real-time holism.

It is not limited to previous familiarity with similar situations.

It involves the capacity to use pieces of knowledge in new ways, i.e. to extrapolate.

The challenge is to find the human way to do that. LLMs, for instance, can extrapolate, but not in a human-like fashion.

Commonsense is an indispensable element of human intelligence.

Situation holism

Situation holism is (roughly) the capacity to coherently articulate two or more ways to render the world intelligible.

It is manifest in the comprehension of simple stories like the Khoja's.

Unlike commonsense, stuation holism is not really worried about the specifics of each kind of situation.

Rather, it accounts for the integration of distinct situations in an intelligible whole.

But just like commonsense, situation holism is poised to play its role whenever we handle the structure of human activities in the world.

For instance, choosing when to enforce a game rule and when to resume real-world inferences requires situation holism.

Situation holism also requires more than mere familiarity with a collection of previous experiences.

We have now a framework-neutral characterization of both commonsense and situation holism.

The non-saturable character of human contexts

The difficulty in explaining commonsense and situation holism is in their reliance on context-sensitivity.

I'll offer here some remarks on what I mean by "context" in general and "human context" in particular.

The set of world's features to which the creature is sensitive constitutes its effective environment.

Efective environments allow creatures to determine which and when to deploy some behavioral output.

Most animals employ a mix of different strategies relying on different levels of environmental stability.

Context is whatever the effective environment becomes under the light of the creature's cognitive capacities.

It is thus relative to the agent's cognitive capacities and extends its effective environment.

We share a good deal of our effective environments with dogs, but the different cognitive capacities mean we can't share the very same context.

This distinctiviness is also manifest in the explanatory roles that cognition and effective environment may play.

What's distinctive about human context is grounded in something distinctive about human cognition.

The context of non-human animals is complex, yet they're always saturated.

Saturated contexts enable the full specification of its features.

And the full specification of how these features are articulated as well.

This is not about levels of complexity: the system's contexts are said saturated when the system can't extrapolate from them in an open-ended fashion.

This is the case in the bird's example, and in the elephant's example as well.

Adding features to saturated contexts does not amount to a change of context.

In contrast, human contexts are non-saturable.

We can get a glimpse of what this amounts to by seeing that even relatively normal and simple situations may involve indefinitely many features.

And indefinitely many articulations of features as well.

Thus, human contexts are not subsets, but a tonality acquired by the agent's whole world (contextual tonality).

The cognitive capacity underpinning non-saturability is a kind of productivity that allows us to extrapolate from any contextual tonality we find ourselves in.

Now we'll explore a bit the explanatory role of context in some well-known cognitive frameworks.

The role of context in cognitive explanations

Context can be not just a mind achievement, but also an explanatory tool to account for the mind's mechanisms..

In order to claim that context has an explanatory role in mechanistic accounts, we must define what we mean by that.

A first possibility is that m-contexts are data structures that help in organizing mental processes.

A second possibility is that m-contexts are a partial causal environment in which we can embed functional specifications.

Both possibilities assume that m-context boundaries can be somehow determined, but we don't know if that can be done.

A m-contextless alternative would be to consider the mind as globally tuned to its environment.

But global state stories don't really buys us anything explanation-wise.

Therefore, context issues cannot ground the choice of any particular explanatory framework.

This is an important upshot, for the role of m-context is usually conflated with the role of representational contents, which is a mistake.

How context sensitivity raises problems about relevance

We have seen that context is the main issue when explaining commonsense and situation holism.

But there's a deeper issue about relevance: how to render the right subset of resources salient for any given task.

What is the connection between relevance and context issues regarding commonsense and situation holism?

Claiming that nothing is ever relevant is obviously a non-starter.

Another option is claiming that every cognitive resource is always relevant, but this turns the issue into a design problem.

That's because we are resource-constrained creatures.

Thus, we're stuck with the need to explain how can one know what's relevant in each and every context.

This introduces a circle: relevance is context relative, but context-sensitivity relies on identifying what's relevant.

To see how hard it is to get out of it, we'll analyse some well-known approaches.

Why relevance problems are so hard

Many strategies to handle the relevance problem (RP) take commonsense for granted.

A path to this error is related to psychology's methodological difficulties in specifying capacities.

To see how far this goes, we'll discuss two deeply different frameworks: Fodor's and Wheeler's.

On Fodor's pessimism

Fodor is famous for defending the language of though hypothesis.

He also claims that the mind's architecture consists of modular and central cognition.

Modules allows for some degree of flexibility, but they're a far cry from what we need for context-sensitivity.

Nonetheless, Fodor insists on encapsulation because he is pessimistic about alternatives such as frames.

He takes the issue to be fatal and claims that no scientific account of central systems is possible.

However, Fodor's fatalism relies on the belief that classical computationalism is all we have.

Furthermore, Fodor thinks that belief revision and scientific theory acceptance are isomorphic processes.

But there are other possibilities: global revision might emerge from local processes, for instance.

Even without this conception of scientific reasoning, Fodor could still say that modules are the only way to avoid RP.

Before digging into the consequences, it must be clear: this is not an issue about computational power.

There are two possible roads to avoid Fodor's pessimism: find non-modular ways to tame central cognition or embrace massive modularity.

Can we dodge Fodor through heuristics?

The first approach appeals to heuristics for fast and frugal processes even in central cognition.

But heuristics fall short of what we need.

They rely on how the information is organized (for instance, minskyan frames).

But only saturated contexts can be organized in this way.

Taking frames to be stereotypical descriptions of contexts which would be further extended thought heuristics doesn't work as well, for there's no such thing as contextless (non-trivial) heuristics.

Heuristics can't guess the information lacking in a frame because they must rely on frames to make those guesses.

If we insist on frames, we end up in a regress where first-order frames rely on second-order frames and so on.

Can we dodge Fodor through massive modularity?

The second road rejects central cognition and embraces massive modularity.

Massively modular systems allow for some flexibility with mutual preemption and complex input clusters.

Sperber uses elements of his Relevance Theory to extend module flexibility in this way.

He relies on s-relevance: a trade off between cognitive benefits and efforts that characterizes inputs.

More specifically, he relies on sensitivity to expected s-relevance in inputs.

But effective calculation of expected s-relevance also relies on something akin to frame systems.

This leads Sperber to claim that expected s-relevance can be handled by physiological, non-cognitive markers.

We can understand such markers as comprising a complex road-like structure routing inter-module interaction.

Sperber and Wilson believe they need not worry about Fodor's construal of RP, for they reject that the mind is Quinean.

But this is not enough to avoid the construal of RP advanced here, so it is a problem for frameworks relying on s-relevance.

The reason is that we need to indefinitely multiply either the modules or the routes among them.

But both approaches raise RP.

Taking stock

The resulting picture suggests the problem is non-computational.

On simplifying things by bringing Heidegger

A non-computational framework in which RP was explicitly addressed is the so called Heideggerian cognitive science.

The connection between Heidegger and cognitive science is not new and can be seen in Dreyfus's early work.

More recently, Wheeler tried to provide a broadly dynamical Heideggerian scientific paradigm for cognitive science.

Wheeler thinks computational systems are a kind of dynamic system, but others think they're different in nature.

The Wheeler-Dreyfus debate on the role of representational contents

The role of computation and representations was the starting point of a debate between Dreyfus and Wheeler.

The debate emphasizes online cognition.

For Dreyfus, there can be no role for representations in online cognition.

But for Wheeler, online cognition can accommodate non-classical representational contents.

These can be conceived as situated representations.

Situated representations block Dreyfus's argument against representations in online cognition.

Dreyfus's point is grounded in a blind spot regarding the possibility of situated content.

The Wheeler-Dreyfus-Rietveld debate on relevance sensitivity

The next topic in the debate is Wheeler's thesis that we can avoid RP by relying on two kinds of couplings.

The first one is the already familiar SPAC (Special-purpose adaptive couplings), and for Wheeler we're essentially a collection of these.

The second one are continuous reciprocal causation (CRC) couplings encompassing brain-body-environment.

Wheeler then analyses RP in two: intra-context (solved by SPACs) and inter-context (solved by CRCs).

The situation is remarkably similar to the one previously found when discussing classical approaches.

Dreyfus rejects Wheeler's analysis and Rietveld also argues against it using a pathology: utilization behavior.

Rietveld claims that UB-patients reveal a primary role for CRCs in explaining relevance-sensitivity.

But Rietveld's argument emphasizes issues regarding SPAC triggering, which present no real problem for Wheeler.

Wheeler agrees with Rietveld that SPAC triggering is the business of CRCs.

Evidence of SPAC-like mechanisms that are resistant to influence from m-contexts is not the issue, either.

Consequently, Rietveld's argument is not enough to devoid SPACs of the explanatory role Wheeler attributes to them.

However, Wheeler's case for SPACs relies on a wanting theory of m-context determination, and the question remains open.

By rejecting the possibility of such a theory, Dreyfus advocates for a framework where mind and world are globally tuned.

His argument is already familiar to us: we need to rely solely on CRCs, and this means to reject any role for m-context.

But the limitation of Dreyfus' suggestion is also familiar: it begs RP instead of throwing light on it.

Wheeler seems to share enough of my view to pressure Dreyfus on the same point.

However, he holds to the distinction between intra-context and inter-context RP and claims that SPACs help with the former, while CRCs throw light on none of them.

But the distinction depends on a theory of m-context, and formulating one is just as challenging as the issue facing CRCs.

Taking non-representational stock

The upshot is that RP cannot ground a migration to dynamic models, for RP is not essentially linked to any framework.

Futhermore, dynamic models are not psychological explanations, for their models comprise psychology's explananda.

Thus, leaving RP aside on grounds of choosing a dynamical framework doesn't amount to a dissolution, but to give up on RP.

What about this nice framework here?

We've seen that computational and a dynamical framework cannot handle commonsense and situation holism for similar reasons.

The reasoning generalizes to all currently available frameworks.

Some ideas might help, but can't solve RP on their own.

Situated cognition is of no help in organizing cognitive resources.

As far as RP goes, it doesn't matter whether information is encoded within our heads or out there.

Emotions also won't help, for they are themselves context-sensitive.

There's also learning, but we can't just learn what's relevant because learning itself presupposes a solution to RP.

The core issue is that general purpose learning is limited.

The problem is specially acute if we need to handle non-saturable contexts.

Mind the gap

Before moving on, there's a worry lurking that must be adressed: what if non-saturability is too strong a requirement for human-like commonsense? Human contexts may be relatively more complex than those involved in the cognition of non-human animals, but perhaps they're equally saturable and the big difference lies in the availability of more memory and processing power.

I think this conflates two distinct capacities, though: the capacity to integrate large amounts of information from distinct sources and the capacity to extrapolate from the results of that integration without losing track of what's relevant. The latter, not the former, is what renders non-saturable contexts.

Though RP may present itself even in creatures without non-saturable contexts, it is the non-saturability that makes a specially hard challenge.

Representational and inferential productivity

In a nutshell, we need to solve RP in order to cope with contexts, commonsense and situation holism.

I suggest we approach RP by means of two aspects: representational productivity and inferential productivity.

Representational productivity is about the mind's representational power.

My suggestion to account for representational productivity is that the mind employs multiple representational architectures.

My suggestion to account for inferential productivity is that we can learn to learn what's relevant, after all.

2 REPRESENTATIONAL PRODUCTIVITY

Handling RP in a representationalist framework forces us to face an issue that does not raise for non-representationalists: how to allow for representational productivity in a non-explosive way. If we manage to do so, we'll have a richer set of tools to account for inferential productivity. In order to show how we can manage that, I take the long road. First, I show in which sense the frame problem from AI is connected (and equally important, in which sense it is not) to RP. I do that by using resources from that literature to delineate the role representations may play. Second, I suggest that the issue can be handled by avoiding sentencial representations and adopting structural representations. It enables us to deal non explicitly with relationships between tokens and contents while providing domain-specific productivity and systematicity. Third, I introduce the concept of representational pluralism and show how the idea of representational redescription allows us to account for world making and its correlative productivity and systematicity in a way that is bottom-up, and not top-down as it is the case with sentencial representations. The resulting picture allows the representationalist to purchase a clear-cut distinction between representational and inferential issues, so she can stop worrying about representations and focus on relevance sensitivity.

2.1 The challenge of representational productivity

Representational productivity is the first *desiderata* outlined at the end of the previous chapter. Though we have already seen that the idea is somehow related to the mind's expressive power, there's a lot of detail to unpack and untangle before we get clear about what's at stake. In cognitive science, the word "productivity" usually means the capacity to have an open-ended set of attitudes. The essential idea is that productive cognitive capacities must be specifiable while abstracting away from time and memory constraints. That's pretty much what we do with computers in general and calculators in particular. We only claim that a calculator can evaluate the product of any two numbers when it is not subject to any constraints other than time and memory. If we could find a huge number that could not be handled solely by managing time (i.e. more time or faster processing) or the available memory, then the calculator's capacity would not be considered open-ended. The most obvious evidence that human cognition involves productive capacities is our ability to understand an open-ended set of linguistic utterances. The productivity of our understanding is one of cognitive science's core *explananda.*

In classical stories, this capacity is accounted for by employing a productive representational scheme. Linguistics taught us the value of compositionality and recursivity in explaining the productivity of natural languages. It allows us to specify an unbounded capacity while abstracting away from time and memory constraints. Why not draw on the same trick when it comes to representations? The most well-known example of this strategy is Fodor's *language of thought* (LOT) (1980). In a LOT-like representational scheme, even a relatively small lexicon gains open-ended expressive power due to the possibility of creating indefinitely many complex compositions. Friends of this approach try to account for representational productivity solely by appeal to the properties of LOT: we have unbounded expressive power because LOT is unbounded. Thus, while the productivity of though is cognitive science's *explananda*, the productivity of LOT is cognitive science's *explanans*.

Unfortunately, the approach brings with it a lot of well-known challenges that are possibly insurmountable. For instance, it is very hard to reconcile with non-human cognition. LOT requires a fundamental architectural assumption about the mind that is not easy to account in evolutionary terms, given its all-or-nothing character. Such a fundamental claim seems to be at odds with the evolutionary history we share with other animals. This is not to deny that there is no gap between human and non-human brains, of course. It is a point about the nature of the gap. Is there some big difference or is it all about scaling up? That's a tricky question, for there are many ways to tell a big difference from a small one. It all depends on which properties of the brain we're interested in attend to. But to claim that the mind is essentially a LOT-powered belief-desire engine has broad consequences for most brain research, which is why it probably qualifies as a huge gap.

Mammalians are a good example: do all of those who have minds employ LOT? Researchers such as Povinelli (2012) provide an increasingly extensive body of evidence about apes' cognitive capacities that makes this hard to believe.[1] Furthermore, one of the core motivations for linguistic representations is explaining our capacity to learn natural languages. But if LOT pervades most (or all) of our closest evolutionary cousins, how come that we're the only mammalian able to do that? On the other hand, to deny an LOT engine for other mammalians amounts to the claim that they have no mind, which is equally odd, specially for those with extensive work in animal cognition, such as De Waal (2016). Many researchers (e.g. Fodor) made an effort to show how psychologically parsimonious LOT is, specially considering the huge challenge of explaining human cognition, but unfortunately it is far from being parsimonious from an evolutionary perspective. Whether there is a plausible evolutionary story to be told remains a largely open question, but given the time and effort such pursuit already took us, with no sign of real success, I stand among those who suspect that's not a good problem to work on.

I don't think friends of LOT will be convinced by these remarks. Many believe that there is no other way to explain the productivity of our thought and understanding. In their view, even though LOT has to face some tough challenges, we'd be far worst without it. We must account for open-ended requirements with finite resources, and that's just what the compositionality akin to that found in natural language seems to buy us. However, I think this is misleading, for it conflates different *explananda*: LOT is able to account for representational productivity, but whether it accounts for inferential productivity is something else entirely. Representational and inferential productivity are both necessary to account for commonsense

[1] I won't dwell on the details now, but Povinelli's research will be addressed in forthcoming discussions.

and situation holism but, taken in isolation, none of them is sufficient. Additionally, I'll argue that the real problem with linguistic representations, is not that they have nothing to say regarding inferential productivity. They're worse than unhelpful, for they prevent us from explaining inferential productivity, pushing us towards a dead end.

LOT allows one to represent an open-ended set of possible situations, as well as their permutations. But they also can represent the same situation in indefinitely many ways and make indefinitely many inferences out of it. To get a glimpse of what's at stake, consider the question "what will happen if I turn this glass of water upside-down?". Now just compare the amount of reasonable possibilities (the water goes down) with the huge amount of logical possibilities (the water goes up, the water changes color, the water stays still, etc.). How can we stop the system from going astray in such a sea of contents that are architecturally possible, yet completely alien to the present circumstances? In order to do so, it must exploit its representational resources efficiently, but in LOT systems, what drives such exploitation are additional representational resources. This raises a snow ball effect where, in order to efficiently exploit a representational resource A one must abide by the rules specified in the representational resource B, and in order to efficiently exploit resource B, one must abide by the rules specified in resource C, and so on indefinitely. This kind of additional instruction is necessary because LOT has no built-in respect for neither the dynamics nor the systematicity of any non-linguistic domain, and therefore can spend lots of valuable time and resources making valid yet completely domain-alien inferences. We need productivity, but we also need to keep it under control. To summarize this view, I'll borrow a term (not the meaning) from philosophy of logic and say that language-like schemes such as LOT are *explosive*.

As we have seen in the previous chapter, the job of taming LOT's explosiveness is all left to mechanisms that try to constrain the set of possible inferences in any given domain (that's what a Minskyan frame is for). But that cannot be done unless we handle RP adequately. However, LOT-powered approaches render RP intractable, for they constrain the set of conceptual tools at hand, leaving the representationalist in a far worse situation when compared to non-representational approaches. In this connection, I regard many of the dead-ends faced by frame systems as products of LOT rather than representations in general. While LOT allows us to at least conceive how finitary can systems deal with open-ended sets of situations, they fall short of enabling an explanation of how finitary systems can make use of compositionality to cope with the world efficiently. I'm not claiming that LOT is the sole responsible for the problem of how to tame cognition, though. We have seen that RP is a problem for non-representationalists as well. The point is that LOT preempts representationalists from exploiting the full potential of their favorite tool (i.e. representations). By adopting LOT, we avoid a system with too little expressive power, but we end up with one we can't keep under control.

In what follows, I'll make an effort to unpack and justify the claims I've just made. I'll ground this pessimism regarding LOT by showing what I take to be the most serious manifestations of its explosiveness: the *frame problem* (FP). The historical pattern of failed attempts to solve it will help us to accept that the issue is likely to be unsolvable. After that, I'll advance a positive suggestion: maybe there is a way to reach out a considerable representational productivity without appealing to explosive schemes. The core idea behind what I'll call *representational pluralism* is to transfer some pressure from inferential issues to the representational schemes themselves and let architectural traits help. However, without a clear comprehension of what makes FP so hard, some motivations behind representational pluralism will look ill-posed. That's why I dedicate the next sections to throw some light on the problem and its relation to language-like schemes.

2.1.1 For the want of a frame

The *frame problem* is a famous, yet sometimes poorly understood, issue for both AI and cognitive science. It was first described in the logicist approach that characterized the work of McCarthy; Hayes (1969). For a while it was extensively discussed within the AI community. In that period, AI was still the intellectual engine of cognitive science, producing both models and hypotheses that could be applied by cognitive scientists in their effort to understand human cognitive capacities. Therefore, the issue quickly got the attention of cognitive scientists like Pylyshyn (1987) and philosophers of mind and psychology like Dennett (1999[1978]) and Fodor (1983). In the following decades, many diagnostics and possible solutions were offered. Some like McDermott (1987) regarded it as a minor issue, while others like Dreyfus; Dreyfus (1987) took it to be unsolvable, and argued that the only way to avoid it is by abandoning representationalism.[2] I meet them half-way: FP is a serious issue, but it is not a product of representations *per se*. We need not abandon representations, but we do need to abandon language-like schemes, at least as far as commonsense and situation holism goes. Given the many interpretations of FP available in the literature, before pointing out how it is connected with the explosiveness of LOT, it is useful to get a bit familiar with the problem and the debate that emerged out of it.

In its first guise, FP was about modelling the effects and non-effects of actions or events. In the 1960s, John McCarthy and Patrick Hayes were developing a logical formalism dubbed *situation calculus* and using it to model artificial agents. The world of the artificial agent comprises a set of axioms describing the current state of affairs at a given time. Call the set describing what's on my table now Ω. The axioms specify the current set of objects, their properties and the relationships they bear to one another. As of now, Ω describes, for instance, the position of the laptop, the coffee cup and the coffee bottle (yep). The properties and relations of these objects are denoted by special variables denominated *fluents*. There can be fluents describing how much coffee there is in the bottle, or how hot it is the coffee that's already in the cup. Fluents might change their values as the situation evolves. The quantity of coffee inside the bottle at situation s1 and situation s2 might be different. The shifting from one situation to another is triggered by actions or events that must be carefully specified. There can be, for

[2] I present an extensive review of that discussion in Barth (2018).

instance, an action dubbed `pour_coffee` to account for the possibility of taking the bottle and use it to pour some coffee in the cup. Therefore, by executing the action `pour_coffee`, the system would change the contents of Ω. Its previous contents described s1, while the new content describes s2. New actions might trigger s3, s4 and so on. Changes in the world are thus accounted by a continuous revision of Ω's contents, and the dynamics of those changes are specified by the set of possible actions and events. To specify what are the consequences of a given action or event is to describe which among Ω's fluents change their values. The action `pour_coffee`, for instance, changes how much coffee there is in the cup (more than before) and in the bottle (less than before).

This leads to an immediate issue about how to model actions and events: which among the set of existing fluents must be affected by any given action? Should `pour_coffee` change fluents such as the one's describing the color of the cup, the final position of the bottle or the quantity of coffee over the keyboard? Sometimes these fluents might change, sometimes not. Once I've put the bottle back on the table, it's position could have changed slightly. If I did not realize how much coffee there already was in the cup, it might overflow and some coffee may spill over the table and reach my keyboard. All of that can also happen to an artificial agent trying to cope with the real world. What McCarthy and Hayes quickly realized is that the choice of fluents could easily get out of control. Can `pour_coffee` change the color of the cup or the bottle? It depends on whether they are transparent or sensitive to some property of the coffee. Can it change the color pattern of the table? Again, it depends whether coffee has been spilled on it. Can it change the shape of the cup? Once more, it depends on the circumstances: if the cup is made of thin plastic and the coffee is too hot, it might happen. That's why we can't simply assume that the action will never have any effect other than a fuller cup.

Therefore, in trying to specify a simple action (pouring some coffee) one finds herself thinking about the possible cases in which that action may or may not have some side effect on the color, shape or other properties of the other objects on the table. But it gets worst: in specifying actions and events, the considered set of objects cannot be comprised of just those that happen to be around in a given situation. One must take into account every existing object type (bottles, cell phones, stones...), and their specific set of properties. We can't limit ourselves to the set of objects currently on the table, as the set itself can change. Actions cannot be bounded to specific situations, for we would need a different action specification for every single situation made possible by Ω. In the resulting picture, to specify what it means to pour some coffee, one has to specify the possible effects of it over every property of every single kind of object that happens to exist in the world of the artificial agent.

In situation calculus, actions are also specified by axioms. The axioms determining the fluents that are always affected somehow by a given action are called *effect axioms*. The axioms that specify which fluents will not be affected by a given action are called *frame axioms*. Frame axioms were situation calculus' tool to set up the boundaries of any given action or event. The problem with frame axioms is that there seems to be no way to rule out any fluent at design time, for the set of affected fluents is context-sensitive and can only be determined at runtime, when the specific circumstances constituting the current situation are already known. Given the need to account for every property of every object that happens to exist in the artificial agent's world, we need a different frame axiom for each available property-action pair. In a world with M possible actions or events and N different properties, the number of necessary frame axioms will be MN. In sufficiently simple worlds, such as the toy or specialized ones we can still find in AI, one can take a deep breath and just write the frame axioms down. However, this is prohibitively large in real world scenarios, specially considering the non-saturable nature of human contexts. The need to avoid an open-ended list of frame axioms is what McCarthy and Hayes dubbed the *frame problem*.

Patrick Hayes took the problem to be essentially that of modeling a principle of *ordinariness* regarding the (agent's) world dynamics. Such a principle would be used as a *ceteris paribus* clause, allowing us to design actions and events in terms like "in ordinary situations, this action has such and such effect and all the rest remains the same". By relying on this principle, we would not need to specify the relationship of every property to everything else in the world just to model the ability to pour some coffee. It would allow us to put all of those frame axioms aside.

So far, FP seems to be a pure design-time problem. There is a practical face and undesired runtime consequences, though. Consider again the situation calculus system Ω, but now with a realistic bunch of knowledge about the human world (i.e. uncountably many frame axioms). Were we to ask it: "if you pour me some coffee, what would happen to the coffee still inside the thermos bottle?". Our answer would probably amount to something like "in ordinary circumstances, and assuming that nothing went wrong, the thermos would have less coffee in it, and that's all". But what would it take for Ω to reach this conclusion? That depends on how the knowledge is stored in the system. Situation calculus' systems could account for such an answer by attending to the event's effect axioms. Thus, if an agent stick to them, there's no risk of losing time in the processing of a huge number of frame axioms. But consider now the question: "if you pour me some coffee, what would happen to it, once it's in my cup?". We know that it's temperature will start to decrease much faster than it was while inside the thermos, and we also know that it's color will remain the same. It's easy to see that the non-change in color should be modeled as a frame axiom. But should an acceleration of the decrease in temperature be modeled as an effect axiom of pouring some coffee? Probably not, for it obviously depends on the cup or whatever is being used as a cup. On the other hand, if it's modeled as a frame axiom, than it will be one that the agent's mechanisms must access and use with reasonable frequency in such situations.

This shows that the distinction between effect and frame axioms may not amount to the distinction between what's ordinary (or typical) and what's not. Consequently, we should not be misled by situations in which what's typical coincides with what's modeled as effect axioms. Frame axioms must be taken into account even in ordinary situations. There is no structural isomorphism between how the information is stored and how likely it is to be used.

With this in mind, we can get back to the question: what would it take for Ω to reach the conclusion that the coffee's temperature will start to decrease much faster than it was while inside the thermos? The answer is that, given the lack of a meaningful structure in the agent's model (as far as ordinariness goes), it would take an exhaustive consideration of all the action's frame axioms. This amounts to pretty much everything Ω's got about the world, for among the processed axioms, there must be things like "pouring some coffee won't change the signal strength of the Wi-Fi networks in range". The upshot is that, despite its expressive power, frame axioms are an incredibly inefficient way to model events or actions. The structure of the stored information should allow for efficient and flexible usage under an open-ended set of possible situations, but the pair effect/frame axioms seems to go in the opposite direction. That's why FP makes very hard to infer efficiently, at runtime, the outcomes of actions and events.

This picture brings with it a serious consequence for (classical, cognitivist) cognitive science. Unless we handle the issue properly, there's no choice but to accept that knowing about X requires knowing about every possible relationship that X could bear with everything else in the (agent's) world. We can see this clearly when considering the capacity to learn about the dynamics of the world. Consider how weird would it be for a system to learn how to execute the action `push_object`.[3] To learn how pushing works goes far beyond getting to know that the object's position changes. It also involves learning that the pushed object will keep its color, that the number of cells in the agent's body won't change, and that the new position has (usually) nothing to do with the importance of math for understanding neural networks. In frame axioms systems, however, all of these ordinarily unrelated matters are considered a necessary part of what it is to know how to push an object. But if one's teaching me what it is to pour some coffee, where does the information that it won't affect my knowledge about the role of neural networks in data science comes from? If the information cannot be found anywhere, what should I put in its place? Hayes' principle of ordinariness seems necessary to acquire knowledge efficiently about any domain.

Just like most hard philosophical problems, FP started as a minor technical issue whose resilience caused it to drawn increasing attention. We can thus learn a lot by getting acquainted with some unsuccessful attempts to solve it. Throughout the 1970s, the issue was largely discussed, and two kinds of strategies to cope with it quickly emerged. Haugeland called them the *cheap test* and the *sleeping dog* approaches (Haugeland, 1987). The crucial idea behind cheap test was to formulate categories of properties according to some criterion and organize the frame axioms accordingly. For instance, one could populate a set with properties regarding relationships between objects (their relative or absolute position in a plane), and another set regarding properties of those objects (mass, shape, color, etc.). Different categories could be then regarded as ordinarily independent of one another. In specifying the set of potential effects comprising some action, one could point out to such categories and

[3] Or equivalently, to learn what are the consequences of the event `object_has_been_pushed`. The issue is not about how to encode know-how.

formulate frame axioms such as "no property of the category x will ever be affected by this". The hope was that it could render the number of frame axioms manageable. The obvious problem, however, is that the world's joints are really hard to catch. Events and actions can affect and change the ground on which the categories are defined. But these were supposed to be the definitions accounting for the world's ordinariness. Patterns of causal connection, for instance, are always subject to revision, precluding the possibility of using them to ground that on which we distinguish effects and non-effects of an action or event. As a result, no one could ever come out with adequate criteria to establish the boundaries of those categories. Indeed, despite the fact that Minsky's frames were not part of the logicist tradition of AI, the same issues these systems faced (largely discussed in the previous chapter) were also faced by logicist cheap test approaches.

As for the sleeping dog strategy, the core idea is to allow a kind of *default reasoning* in which, as the name suggests, only a reasonable subset of the axioms gets considered by default.[4] Everything else is taken to be a non-effect and is ordinarily ignored. Thus, unless the system has good-enough motivation to go beyond its default reasoning, it will stick to it. The problem, however, should be already clear: how can the system know whether it has good-enough motivation to consider some non-default effect? The FP is now the problem of categorizing the world in facts that can be (ordinarily) ignored and facts that can't. Note that this is not an epistemic issue in the sense that there must be a way for the agent to get things absolutely right. All that is needed is a way for the agent to guess above chance what a good-enough motivation is. Some researchers claimed that this can be done quite easily (Lormand, 1996; McDermott, 1987). However, this is true only of artificially saturated contexts. Indeed, in simple enough saturated contexts, that would be true of frame axioms systems as well. Writing all those axioms would be tedious, but not unmanageable, leading one to claim that the solution to FP is to get an intern.

Despite issues like these, the sleeping dog strategy was considered quite appealing among AI researchers. It triggered the development of non-monotonic logical formalisms such as those we find in McDermott; Doyle (1980), McCarthy (1980) and Shanahan (1997).[5] The underlying assumption was that FP is a product of the monotonicity of first order logic. Assuming that Ω is monotonic, adding new axioms to it will allow new inferences to be made (which amounts to changing the current situation), but will not preempt the system from deriving what could already be inferred in the previous situation. This means that sometimes Ω will allow the system to infer things that were true a long time ago (such as an object's position). One of the reasons why situations calculus needed frame axioms was to avoid this kind of inconsistency. Frame axioms were supposed to narrow down the system's inferential capacities to those allowed by the current situation. *Prima facie*, non-monotonic formalisms might seem useful to overcome this without the need of frame axioms. But this is misleading. They fare no better than any other sleeping dog approach, for they don't avoid the need to

[4] The name comes from the saying "let sleeping dogs lie".

[5] It is also behind the approach of Shanahan (1997).

know when the system is supposed to stick to its default reasoning and when it's not. All we have is a different formalism in which some eventual solution could be expressed. We remain clueless as to what could be the contents of anything like a principle of ordinariness.

What are we to do of all this? We can clearly see that FP is closely related to the issues discussed in the previous chapter, but what exactly is the nature of this relationship? Throughout the years, this has been extensively debated (Ford; Pylyshyn, 1996; Kiverstein; Wheeler, 2012; Pylyshyn, 1987) and distinct approaches were put forward. In order to ground my claim that FP is a manifestation of the explosiveness of a scheme, we have to rule other interpretations out. In particular, we have to discuss its relationship with RP.

2.1.2 What is the frame problem about?

The chapter started with the claim that FP is a problem about representational productivity related to the explosiveness of LOT. This claim is yet to be warranted, though, for the contemporary canonical interpretation of FP is that it is identical to RP. For instance, Fodor and Wheeler, whose work was discussed in the previous chapter, took RP to be the referent of the expression "frame problem".[6] Therefore, before developing our diagnostic further, it is worth dedicating a bit of time to understand why should we accept that FP an RP are not identical, despite being closely connected in representational frameworks.

Dennett was probably the first philosopher to realize that the FP has important philosophical implications (Dennett, 1981, 1987). But in doing so, he took the problem to be about relevance: FP should be regarded as an instance of RP. His first example was about belief fixation: given a new piece of knowledge, how should it affect what the agent already knows? Only the relevant portions of the agent's knowledge should change, but how do we identify these, i.e. how do we frame them? If the system knows what is relevant in the current circumstances, all those frame axioms will not be a problem, for it will ignore the irrelevant ones. In this view, the difficulty to find categories for cheap tests or criteria for default reasoning is pretty much the same we face when trying to account for what is relevant in non-saturable human contexts. The original formulation of the problem is thus regarded as the way RP manifests itself in situation calculus. Consequently, as far the need for a *ceteris paribus* clause goes, a theory of relevance would amount to a theory of ordinariness and vice versa. Dennett's reading became influential in the philosophical community. Throughout the 1970s and the 1980s this tendency (mostly among philosophers) to regard the problem of handling frame axioms in situation calculus as a broader issue caused the eruption of a "terminology war" about what should be the referent of the expression "frame problem" (Hayes, 1987). Despite some complaints from the AI community, the semantic battle was lost and nowadays most researchers in philosophy of cognition understand FP as RP.

Dennett's move is an extrapolation of the practical issue outlined in the previous section: the unrealistic need to continuously process every single frame axiom even in ordinary

[6] This view of FP as an instance of RP was also my view in Barth (2018).

situations. There are at least two significant problems with this move, though. First, it conflates issues of representational and inferential productivity. It tries to alleviate or dismiss representational issues by treating them as an inferential problem. But as we have seen, representational productivity is about unbounded expressive power, while inferential productivity is about being able to exploit one's cognitive resources (including one's expressive power) in an open-ended set of circumstances. The temptation to conflate these issues comes from the explosiveness of LOT. Since the scheme itself puts no constraints on what can be inferred, the system is architecturally free to derive indefinitely many permutations, even those alien to its current interests and expectations. Likewise, it is free to theorize (i.e. model) the same situation in multiple ways, which creates the problem of selecting the appropriate theory. Therefore, once we accept LOT, the only way to constrain its representational power is by taming the system's inferential capacities. That's what cheap tests and sleeping dogs are all about, and we've just seen why they fail. The history of failed attempts to handle FP as an issue about inferential productivity suggests that the approach leads us to a dead end.

The second problem with Dennett's move is that it conflates issues about ordinariness with issues about relevance. Handling relevance is distinct from handling ordinariness. The way we structure and articulate the stored cognitive knowledge to account for relevance may be very different from the way we do so in articulating ordinariness. Consider the classical formulation of FP as an issue about modelling non-effects of actions and events. If some property is not affected in any way by some action, that doesn't mean it is irrelevant. There are many circumstances in which a non-effect can be actually relevant ("See? Nothing happens when I put it there!"). Thus, even if we could find a solution to the FP in classical approaches (such as situation calculus), the question of how do we tell what's relevant from what's not, would remain wide open.

What about the converse? Can a theory of relevance work as a theory of ordinariness? We can surely conceive some artificially constrained contexts (akin to GOFAI's micro-worlds) in which these might conflate. A specialized system or a very simple creature certainly could store its cognitive knowledge about the world in a way that gives it exactly what it needs whenever it needs. This means that relevance and ordinariness can look alike, but only within saturated contexts. In more realistic cases, though, that's not an option. By analogy, a researcher relying on lots of reading notes to keep its work going knows how to discern the relevant ones according to its current needs. Yet, if the notes are not stored in a way that allows for efficient retrieval, she will waste time until she finds what's needed. If she has a well-organized set of notes, she'll know that knowledge about chess championships and dogs usually don't go together, and yet, if some circumstance in which both are necessary arises, she'll be able to retrieve that cognitive knowledge quickly. Thus, even if we could formulate some principle of relevance (on grounds of which she realizes the need to use what she knows about dogs and chess), we would still have to face the problem of how to store cognitive knowledge in a way that allows for efficient retrieval. Whatever inferential strategies we used to cope with relevance must rely on this. In the same vein, ordinariness is about expressing and storing the agent's world dynamics in a way that allows for efficient retrieval and exploitation, whatever the circumstances. The upshot is that FP and RP have distinct *explananda*.

I'm not alone in taking FP to be distinct from RP. Lormand (1996), for instance, follows Hayes (1987) and rejects Dennett's account. Their point is not that philosophers like Dennett fail to capture a real issue about relevance, but only that they're capturing something distinct from FP. Lormand claims that the FP is a very narrow issue about logically modelling the *persistence* we find in the world's dynamics. If we can get the system to work without lots of frame axioms, then the issue is solved, and that's exactly what the sleeping dog approach allows us to do. Thus, even if we claim that the sleeping dog approach won't help with non-saturable contexts, Lormand would insist that this is outside the approach's scope. In a nutshell, philosophers like Dennett would be guilty of a big mess, for their rejection of the sleeping dog strategy is based on its ineffectiveness against RP. But the target was never RP, only FP. By analogy, Lormand's point is that philosophers are rejecting headache pills due to its ineffectiveness against cancer.

I take FP to be the manifestation of an issue about representational productivity, so I obviously agree that FP and RP are distinct. However, *pace* Lormand, I don't think FP can be solved by appealing to sleeping dog approaches. The deeper reasons for this claim will be made clear throughout the chapter, but for now it suffices to say that if we stick to Lormand's reading of the issue, then the sleeping dog approach won't buy us anything. Here's why: Even in its very first formulation (McCarthy; Hayes, 1969), FP was conceived as an issue that preempts the escalation of situation calculus' models to real-world human circumstances, i.e. to non-saturable contexts. The core issue with frame axioms was thus that only saturable contexts (such as those we find in classical AI's micro-worlds) could be dealt with. If we're willing to give up on more realistic contexts, then there're all kinds of non-interesting solutions available. For instance, we could just provide a single rule according to which nothing changes unless whatever has been made explicit in effect axioms. Strategies like these are the reason we can play with chess engines in our computers.[7] Though one could argue that, in a very loose sense, this would amount to a version of sleeping dog, the point is that this approach is entirely compatible with situation calculus. Thus, all the effort put in developing, e.g. non-monotonic formalisms, didn't actually buy us anything helpful regarding FP. If we give up on modeling non-saturable contexts, FP becomes just a minor technical nuisance. In that case, the most appropriate answer would be to agree with McDermott (1987) and question the very existence of a serious problem. We end up in a seesaw: if we insist in handling non-saturable contexts, the sleeping dog approach fails to deliver dividends. But if we stick to saturated contexts, then the sleeping dog delivers nothing we couldn't already buy with frame axioms as well.

The discussion of the past attempts to solve or dodge FP is helpful in getting a clear picture of what's at stake. Despite the disagreements about some details, I believe that the

[7] Just to be clear, I'm not claiming that current chess engines are designed with situation calculus. The point is that they use the principle according to which only explicitly mapped consequences are taken into consideration.

interpretation of the FP as an issue about representational productivity is reasonably faithful to the original FP conception. It takes the problem to be about *modeling* change, that is, about allowing suitable representations of the possible changes in the agent's world through time. On the one hand, by insisting on the need to handle ordinariness in non-saturable contexts, we avoid portraying the issue as less severe than it really is, and this leads to the rejection of strategies that can't really buy us what's needed. On the other, by rejecting the conflation of the issue with RP, we avoid the rejection of possible solutions (or dissolutions) due to traits pertaining only to RP. Handling FP is a necessary step for any representational framework targeting RP, though it's far from sufficient. It is necessary because FP is about keeping representational resources manageable, and no real solution to RP is possible unless we manage to solve or dodge FP.

Lets quickly take stock. The upshot of the previous discussion is that FP is a manifestation of LOT's explosiveness. Whenever we apply linguistic schemes in modeling the world's physics, the structure of a song or linguistic utterances, FP is bound to appear. When modelling the dynamics of such domains, one must make all the possible permutations explicit in the form of rules. However, there is a huge gap between the scheme's expressive power and the set of permutations allowed by any given domain. Most of the permutations that can be logically expressed are alien to that domain and must be precluded. Consider a causal domain where objects usually go down. LOT allows for objects going in every logically conceivable direction. Consequently, not just every possible permutation allowed by a domain must be explicitly stated: in order to avoid odd behavior, all the abnormal ones must be included in the model as well. But the only way to constrain its representational power is by adding more rules, i.e. by storing more cognitive knowledge about what inferential paths it should take when using that cognitive knowledge. The more rules are added, however, the more rules need to be further added so that the system knows how to cope with all of them. In systems capable of handling non-saturable contexts, this implies an unbounded number of additional data-structures, which renders the task hopeless. That's why, in a nutshell, any LOT system trying to cope with non-saturable contexts will manifest FP.

Considering the diagnostic, one can't help but wonder whether it's worth to stick with LOT. Indeed, I'll claim that getting rid of LOT in accounts of the mind's representational contents does allow us to avoid FP. Unfortunately, the cost is huge, for now we have to give up all of LOT's purchases, including representational productivity. As we can't simply ignore the need for representational productivity, if we want to avoid language-like representations, we'd better come up with something else to fill this gap. Of course, that's easier said than done, and considering how drastic the suggestion is, my approach will be to challenge the LOT tradition constructively by outlining an alternative dubbed *representational pluralism.*

2.2 Structural representations and the frame problem

Before digging into the main point, there are important preliminary remarks to keep in mind. In what follows, I'll discuss some properties of representations without worrying (much) about their psychological plausibility or implementation details. Whether it's possible to account for a given representation mechanistically, or whether such candidate mechanism can be found in the brain are questions that will be postponed to the next chapter. Foundational issues will be put aside as well. Thus, the question of how a given physical state (presumably neural) can bear contents will not be discussed at this point. This is why the examples I'll use vary somewhat freely even between personal and sub-personal scenarios. For instance, when saying that a representation can be exploited by its user in a certain way, the term "user" does not imply a fully-fledged agent. Sub-personal mechanisms or AI systems may be regarded as "users" of representations as well. This move is purely methodological. I have no intention to put forward anything remotely close to the classical claim that cognition comprises an independent domain of inquiry. The wetware in which (as well as the body through which) some capacity is exercised definitely has a say in accounts of that capacity. However, when discussing the role of representations in cognition, it is important to be clear about what's a feature of the content being exploited and what's a feature of something else. That allows one to establish a clear boundary between representational and inferential assets, specially regarding their respective senses of productivity.

But how are we suppose to understand representations at such an abstract level? In order to avoid *ad hoc* stipulations of what should be regarded representational, I'll use Haugeland's classical account as a starting point (Haugeland, 1998e). The core idea is that representations can work as stand-ins for something else. For instance, whenever a pre-generation-Z uses a physical map to walk around a city, the map allows what Swoyer (1991) dubbed *surrogative reasoning* about the city while thinking. The idea that maps are representations because they stand in for something is intuitively palatable, but there are other kinds of stand-ins that are harder to bite. Gastric juice is a well-know example. It can help a predator to track its prey even in the absence of a direct causal connection, for its presence preempts the organism from going astray in case it momentarily looses the prey's scent. In this sense, it can be said to stand in for the animal's scent. Stories like these are usually presented as unwanted consequences of naive representational theories, particularly those relying on covariational information. The current need is peculiar, though. While such cases have to be ruled out, it has to be done in a way that does not imply commitment to any particular theory of representational content. Haugeland manages to do it by deeming as representational only the stand-ins whose content (i.e. that which they stand in for) is determined by virtue of an encompassing *representational scheme*. To be so regarded, a representational scheme must satisfy the following:

> (...) (i) a variety of possible contents can be represented by a corresponding variety of possible representations; (ii) what any given representation (item, pattern, state, event, ...) represents is determined by some consistent or systematic way by the scheme; and (iii) there are proper (and improper) ways

> of producing, maintaining, modifying, and/or using the various representations under various environmental and other conditions. (Haugeland, 1998e, p. 198)

The most intuitive example of a scheme that satisfies Haugeland's criteria is a formal language. Roughly, for any linguistic token there's a corresponding linguistic content, and the correspondence is established systematically, so one has to abide by the formalism's rules in order to have well-formed formulas. However, Haugeland's characterization is liberal enough to accommodate both language-like schemes and structural schemes (maps, graphs, geometric palettes, etc.).

Structural and linguistic schemes are classically distinguished by the nature of the relationship between tokens and contents.[8] In linguistic representations, the scheme is responsible for allowing a given primitive term to partake in complex well-formed formulas. But at least in standard stories, in claiming that a given term means x, what one is really saying is that the term refers to x. Thus, primitive symbols of linguistic schemes are said to represent by reference. The term "reference", as used here, must be understood in a loose sense. It captures the idea that the contents of the scheme's primitives are detached from their formal properties. For instance, the word "dog" refers to dogs in virtue of something alien to the scheme, such as a convention.[9] However, the scheme's formal properties are crucial to determine the contents of the compositions and articulations they allow, i.e. the meaning of sentences is a function of their primitive constituents and the scheme's syntactic rules.

In contrast, structural representations get their contents by being structurally isomorphic to them: a given map is said to represent city streets in virtue of the structures they share. This does not imply, of course, anything like visual resemblance. Though that is a kind of isomorphism, the structures in question are usually highly abstract. In this sense, a graph may be isomorphic to the causal structure of some domain. Consider how distinct this is from relying on reference: just to begin with, whether a given structure represents a dog is all about its formal properties. There's no space for references in defining the role played by the elements of a structure. Of course, you can use structures as symbols that refer to something else. A dog-structure can be used to refer to an abstract concept such as friendship. But the structure won't stop being isomorphic to the dog just because you're not exploiting its structural properties. Furthermore, the word "isomorphism" must be taken in a less strict sense than the one usually employed by mathematicians. Most applications of the idea make room for less-than-perfect mappings between structures. If we were too strict, there would be no room for the statistical analysis typically used to compare them (consider how we use fingerprints in identification tasks). This buy us the possibility of degrees of accuracy, i.e. we can talk about

[8] The grounds on which the classical distinction rests are fundamentally flawed, though. In order to properly ground the pluralism about representational schemes, It'll be challenged and replaced. But the replacement bears no weight on the current discussion.

[9] I use this example for the sake of simplicity. I'm not assuming that linguistic schemes can account for natural languages.

how isomorphic a structure is to another one.[10]

With this distinction in mind, it's time to dig further on the purchases allowed by structural representations.

2.2.1 What structural representations buy us

Can structural representations buy us something that language-like schemes cannot? In what follows, I'll argue that the answer is yes. They enable the system to exploit the systematic permutations allowed by a scheme without making room for FP. In other words, they're not necessarily explosive. Structural representations do not require an open-ended amount of additional representational assets in order to be adequately exploited. One can see this feature at work in the case of changes in the world that require changes in its representation, which is precisely where FP bites. This is synthesized in Erik Janlert's requirement of representational conservativeness:

> A sign that the frame problem is under proper control is that the representation can be incrementally extended: a conservative addition to the furniture of the world would involve only a conservative addition to the representation. (Janlert, 1996, p. 40)

Consider again the representation of whatever is in my table (Ω) and let's say that at some point a new cup of coffee has been added. How is this to be handled? Linguistic representations fail to satisfy Janlert's requirement because even conservative changes yield an explosive need for additional representational assets. We've seen that happening with situation calculus. Let's quickly recapitulate how it goes. In order to render a proper model of Ω, we need to describe the properties of every object (color, size, shape...) the possible actions (push, pull, change place...), as well as the relationship each object bears to one another, the relationship between its own properties and the properties of other objects, the relationship between every property and every outcome of every possible action, and so on. Whenever a new object (or property, or action) is introduced, we need to make sure that the model remains accurate and useful. Otherwise, it won't be able to guide the system in exploiting the situation, i.e. the system can't rely on them to make plans and predict outcomes. In order account for the new cup, we would need to add rules mapping the object's properties and relationships to all the other already properties and relationships, as well as every possible action or event. Thus, even a conservative addition (it's just another cup) to whatever is being modeled requires a potentially explosive number of additional representational assets.

Nothing similar is necessary whenever we employ structural representations. A good example can be found in Haugeland (1987). Consider the map-like depiction of the spatial re-

[10] In order to emphasize this point (and avoid confusion with math departments), some authors prefer to use the word "homomorphism". I have no quarrel with that, but I don't think it really buys us anything. What matters is the underlying mathematical account of accuracy. An appropriate account might be equally incompatible with what mathematicians understand by homomorphism as well.

lationships between A, B and C in figure 1, and the following set of entries in an LOT-powered system: [11]

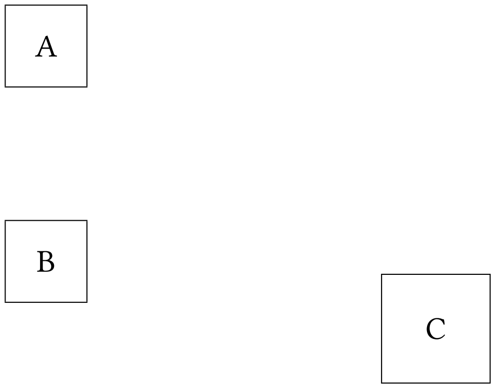

Figure 1 – A-B-C spatial relationships.

(1) A is 10 miles north of B

(2) A is 20 miles northwest of C

Both the language-like entries and the structural (picture-like) map account for the relationship between B and C. However, in the linguistic case the B-C relation is implicit, that is, it must be inferred through some kind of reasoning process over the conjunction of (1) and (2). Consequently, if we change the contents of either (1) or (2), then the B-C relationship is going to "automatically" change with it as well, for now the conjunction of (1) and (2) yields a distinct set of valid inferences. But this is not at all the case with the A-C relationship. Given that it is explicitly stated in (2), it must always be explicitly updated as well. For instance, say the system gets the following input.

(3) A is 30 miles northwest of C

Once the system accommodates (3), its state will be inconsistent, for (3) contradicts (2). The inconsistency is solvable, but the crucial point is that the representations won't do a thing to help with that. It is entirely up to the user (i.e. the system exploiting them) to detect and figure out a plausible way to discipline and keep the set of entries consistent and useful. This is odd, for representations were the entities supposed to explain how a system can accomplish this kind of task. In a truly representational account, the system's reasoning should be a product of the available representational resources, not the other way around. We already know how linguistic systems try to keep this explanatory role for representations: by adding more and more explicit entries aimed at disciplining the system's reasoning and precluding idle inferences. Unfortunately, this is where FP becomes salient.

Contrast this with the case of the picture-like map. Should we say that the B-C relationship is explicit in the picture? In a sense, sure. The relationship is there for us to see. But if

[11] Figure 1 is an adaptation of the one found in Haugeland (1987).

we change the position of, say, A, then all relationships with both B and C would be "updated" just like what's implicit in linguistic entries. The information about B-C is explicitly there, but it works just as if it was implicit. Importantly, whether the position A-C is implicit or explicit in the map depiction is a question that can't even be formulated, for unlike the linguistic construal, there is no sense in which the relationship can be regarded as derivative of the position of the other points. They're all equally primordial. This is a common trait among structural representations, so it seems that the possibility of an implicit/explicit distinction is an artifact of the chosen representational kind. Structural representations rely on schemes that preclude the distinction.

This somewhat odd conflation of implicitness and explicitness means that, when adding a new feature, we don't need to add an explicit set of additional rules in order to keep the representation useful. Consider what would happen if we added a new element D somewhere in the map-like structural depiction (figure 1). All of its geometrical relationships with the other elements would be already established, i.e. we'd be able to know whether D is larger than A, or what is its relative position. There is no need for further information, nor there is need for some kind of reasoning in order to verify the representation's consistency. Right after the entrance of the new element, the representation continues to be an effective guide to the system's reasoning and behavior. This is a way to satisfy Janlert's constraint of conservativeness and avoid explosiveness.

2.2.1.1 Representing change

Structural representations enable an additional purchase that will prove very valuable in our forthcoming attempt to handle RP. Consider *folk physics*: the amount of things we know about the physical world without taking classes. Things such as: objects tend to fall; towels within a water tank will bring some of that water with them if removed; sitting on eggs usually break them, and so on. Folk physics is a recurrent theme within the FP literature, for many of its most typical examples are about causal domains. Will changing a tea cup's position also affect the position of the plate in which it was on? Usually no. Will changing the position of a plate also affect the tea cup that's on it? Usually yes. Indeed, the emphasis on constructions of FP within causal domains was such, that some took it to be an issue about formulating an adequate theory for folk physics (Glymour, 1987). As we have seen when characterizing commonsense and situation holism, FP is definitely not limited to causal domains, but these still make for good examples.

Structural representations are specially useful to target causal domains. That's not just because they can capture contents in a way that avoid the intrinsic/extrinsic distinction. Its contents enable a distinct way to capture the world's features, and consequently allows for distinctive explanatory strategies of cognitive capacities. As we have been discussing, FP is about handling the dynamics of the contexts faced by the system, i.e. it is about modelling change. The dynamic nature of the world was the Achilles' heels of situation calculus and other language-like schemes. In those systems, remember, situation models (or theories) are sets of clear cut snapshots and rules (laws) governing the transition from one state to the next. The world is divided in clear-cut situations describing the current states of affairs, and the dynamics of each underlying domain is expressed in sets of rules used to discipline the representations. Therefore, it's up to the system's exploitation processes to effectively apply those rules and calculate the transition effects while keeping the representations consistent and under control, somewhat like Laplace would imagine God's work. Unfortunately, as we have seen, trying to model such rules using language-like schemes raises FP.

In contrast, by using structural representations we can purchase an additional strategy: rather than formulating rules to tame the user's exploitation of the representational resources, we can actually represent the dynamics. The change itself can be pictured, i.e. represented rather than calculated. This is how Cummins summarizes this point:

> Instead of changing your representations, you can represent change. Instead of disciplining the representations in a way that mirrors the dynamics you are interested in, you can put the dynamic structure in the representation, that is, represent the dynamics. (Cummins, 1996, p. 95)

In doing so, representations can get their explanatory burden back: they can guide the system in handling tasks that rely on the dynamics of the target domain. The burden of the system's reasoning processes can be thus alleviated. It can let itself be causally guided by its own representations, just like it can let itself be causally guided by the environment. In this sense, the system can take advantage of the structural representations to think about causal domains just like we take advantage of scale models when thinking ahead and predicting outcomes.

The strategy allows us to overcome the need to carve out the world as sets of time-instants. We can shape the trajectory of any given set of properties or entities throughout an n-dimensional space comprising a domain. In other words, we can picture objects falling or transforming themselves through time. Instead of time-instants, such trajectories comprises intervals or "stories" within the domain. Such stories can themselves become the building blocks to be exploited in reasoning processes, as well as their possible systematic permutations. The system can use them just like we do when comparing distinct paths in maps or distinct simulations in scale-models.

Furthermore, the involved time intervals need not be short. Contemporary AI systems promising to show how our faces will look when older rely on this kind of pictured trajectories. This strategy is not available to language-like schemes because the carving out of those intervals also rely on the absence of an implicit/explicit distinction.[12] When the distinction is not preempted, the system is forced back to the need for disciplining its own representations and the capacity to satisfy Janlert's constraint is lost. In particular, we'll face the problem of where to draw the boundaries between any two stories, which is simply another guise for FP:

[12] In more precise terms, it relies on architectural features of structural schemes. This point will be further developed in a bit.

the issue of how do we model what changes (and what doesn't) between any two time instants is now the issue of finding out what changes between any two time intervals. To be sure, the possibility of representing change in this way doesn't mean that we should get rid of the old approach in which the system is responsible for disciplining the representations. That is an important strategy, and sometimes it's all we have. But structural representations purchase an important and fundamental explanatory approach (at least regarding FP).

2.2.1.2 Intrinsic contents in structural representations

The suggestion that structural representations can get us rid of FP is not new. It can be already be found in Haugeland (1987). Why was it thoroughly ignored for so long? I think there's multiple reasons for that. Some have directly rejected the idea that structural representations can avoid FP (Hendricks, 2006), but their arguments depended on taking it to be an instance of RP, which is false.[13] Others worried about issues that are unspecific to either FP or RP, such as the apparently diminished role that structures can play in accounting for attitude contents (what could be the structure underpinning the belief that democracy is good?). These kinds of concern will be properly addressed in chapter 3. But if the reliance on structural representations to handle FP is to get off the ground, some concerns require immediate attention. What is it about structural schemes that allow them to do that? We need to understand a bit more about the nature of this feature in order to know e.g. whether it is possessed by all and every structural scheme or only some. Furthermore, we're yet to see how exactly structural representations can account for representational productivity, and we have to make sure that the necessary features are still at work even on more realistic and demanding cases. If the price to handle FP is to be paid in expressive power money, then the cost is too high, for we can't go on without representational productivity.

We can find an important clue in Palmer (1978): the heart of the matter is that structural representations can use their own relational properties as stand-ins for the relationships among whatever is being represented. In other words, they can *intrinsically* mirror the dynamics and systematicity of whatever they represent. A relationship such as "C is larger than B" need not be explicitly stated in some additional token, for the information is already there in the map-like depiction to be exploited (check back the figure 1 if unsure). Thus, whenever one changes the geometrical properties of some object in a map-like structure (location, size, shape and so on) its geometric relationship to everything else is "automatically" updated. In particular, the A-B-C map-like example shows how we can reason about geometric relationships in the world by using the geometrical relationships of a representation as a proxy. That's an instance of surrogative reasoning. Consequently, inasmuch as the geometrical properties of a token accurately map the structure of its wordly counterpart, Janlert's conservatism requirement can be satisfied and the allowed systematic permutations can be expressed without the risk of a representational explosion.

[13] Recall the recent discussion on the nature of FP.

But is the preclusion of the explicit/implicit distinction a general feature of any structural scheme or is it a special feature peculiar to some of them? Consider what happens if we use a 2-dimensional scheme (plane) in order to model a 3-dimensional domain (space). The extra dimension comprising space allows for relationships such as "x is on top of y". Could we try to account for the missing dimension using a 2D scheme? Sure. Topographical maps are an example of how we can do that. However, the relationship "on top of" is not captured intrinsically by the 2D scheme, and consequently it must be disciplined by the system employing it. For instance, topographical maps must be evaluated by assuming a view point from above, but this information is nowhere in the token. It's up to the scheme's user (i.e. the system exploiting representations forged with it) to discern whether the containment of any geometric shape within another is to be regarded as an "on top of" relationship. It could also consider this an inconsistent state in which two objects occupy the same space at the same time, for instance. This is a crucial difference, for the representation cannot play its guiding role in any matter regarding the missing dimension. We're in a scenario where the system is responsible for keeping that aspect of the representation accurate and useful, and that seems dangerously close to what we face when using linguistic schemes.

None of that would be necessary with a 3D scheme, for it already encompasses the missing dimension. Therefore, a 3D scheme is able to avoid the explicit/implicit distinction in 3D domains, and a 2D scheme is able to avoid it in 2D domains, but the 2D scheme cannot fully avoid it in 3D domains. This can be generalized, so we find the answer to our question: intrinsicness relies on the resemblance between a structured scheme and a structured domain. Both have a bunch of dimensions, and a set of basic elements that can be used to fill them up. Rather than points, the palette of basic elements may comprise complex forms (squares, triangles, circles, and so on). Inasmuch as they share the dimensions and the set of basic elements, scheme and domain also share the set of possible distributions of their basic elements alongside the available dimensions. In this sense, every structural scheme can be said to have a target domain. They should always be considered specialized and domain-specific.

The upshot is that intrinsicness is a product of specialized schemes, i.e. tailor-made architectural features. As long as they're only employed within their target domain, structural schemes enable the system to forge representations that can work as effective guides without raising FP. The scheme itself preempts the system from going astray, for no permutation alien to the domain in question is architecturally possible. This also means, among other things, that there is a wider explanatory role for architectural traits than representationalists themselves usually acknowledge.

2.2.2 The explanatory role of architectural constraints

A scheme can be part of the explanation of how a system can efficiently cope with a given domain through representational means. They can constrain both how the content can be processed and what can be represented. At the agential level, this is hardly news.

Consider the need to calculate the product of any two numbers. When the number domain is expressed and handled through the Arabic numeric system (1, 2, 3...), one can use the partial products algorithm that we all learn in school. The algorithm exploits structural aspects of the domain that are adequately captured by the scheme (notice the importance of placing the partial products in the right position before summing them up). These aspects are not available in other schemes, such as the Roman system (I, II, III...). Its hypothetical users would have to make successive sums, which is often less efficient and more time-consuming. Furthermore, the situation is the same we have found when comparing 2D and 3D schemes: in the absence of the structural features, the user has to rely more heavily on its own skills. The Roman system fails to capture a significant structural aspect of the number domain. That's why it can play no role in the explanation of the user's capacity to multiply. The scheme becomes a tool used to merely express the results she's coming up with by other means.

Architectural traits (i.e. features of the representational schemes) can play similar explanatory roles at the sub-agential level. When a scheme fails to capture a systematic aspect of some domain, its tokens won't be able to help the user in handling that aspect. If the system can't or refuses to take care of that, it becomes subject to what Cummins dubbed *forced error* (Cummins, 1996). Whenever one tries to represent content beyond the scheme's reach, the token will be necessarily inaccurate. It is well-known, for instance, that the Euclidean geometry is unable to capture in full the structure of physical space. But there are many other examples: three-colored schemes such as RGB (for red-green-blue), who can be found in most image editing software, can't represent transparency. If one tries to token what would be an RGB-powered representation of transparency, one ends up representing some shade of color instead. Likewise, musical notation can't express accurately a speech or a receding train whistle, city maps can't represent distances with precision - for this also depends on how hilly the terrain is - and so on.

Forced errors are a kind of expressive constraint architecturally imposed on the system. They are different from other kinds of representational errors because the system can't help itself. This possibility should come as no surprise for anyone willing to accept the existence of domain-targeted schemes. After all, it would be very difficult to build an accurate topographical map out of the shapes used in electrical diagrams. To see the difference more clearly, consider these examples of "unforced" errors: a visual system may represent an animal or object as larger than it really its. Someone might take a cow for a horse. A student might represent "I won't do it unless she does" as $S \supset I$ rather than $I \supset S$. All of these are "unforced" because in none of them the adequate content is beyond the expressive power of the employed scheme. In such cases, the error is due to the system's limitations in the capacity to build representations with that scheme. The upshot is that claims regarding a system's capacity to exploit some representational scheme are distinct from claims about the set of contents allowed by that scheme. While the latter is a claim about architecturally imposed constraints on the system's expressive power, the former is about how well the system is able to exploit the expressive power architecturally afforded. Thus, if a system can fully exploit a single representational scheme, its expressive power is as big as the scheme's. But if the system can't exploit that scheme's expressive power in full, its overall expressive capacity gets decreased.

With this in mind, Cummins suggests that such content-constraining properties of a scheme should be interpreted as *a priori* assumptions about its target domain (Cummins, 2010b). In this sense, domain-specific assumptions like "every object has color, shape and size", "no two objects can occupy the same place at the same time" or "objects remain still unless forced" can all be architecturally established. Cummins presents this point with the example of an artificial structural scheme called *Cubic*.[14] Cubic is a kind of gigantic 3D Rubiks's cube. Each of its dimensions comprises 10,000 smaller 3D cubes whose surface can be colored. Cubic can thus represent various coloring patterns in a huge 3D space. A perceptual system could use Cubic to represent not only objects as viewed from the outside, but also their internal structure. For instance, a token whose contents were a bottle full of coffee would be distinct from a token representing the same bottle, but empty. What's interesting about this scheme is that it can architecturally ground the following (very familiar) *a priori* assumptions about its target domain:

(1) Every object is colored.
(2) Every object has a determinate size and shape.
(3) No two objects can occupy the same place at the same time.
(4) Every object has a determinate location relative to every other object.
(5) Every object is at a determinate distance from every other object.[15]

None of those assumptions is explicitly stated anywhere in Cubic. Nonetheless, they're all enforced by the scheme: if a user (e.g. some perceptual system) tries to represent a counterexample to any of those assumptions, the result is forced error. For instance, it simply can't represent an object's size without representing its shape as well. The same assumptions can be regarded as axiomatic while using Cubic to reason about a domain. In this sense, it's user can rely on this set of assumptions without the need to render them explicit or to enforce them somehow: the scheme itself takes care of that.

Whether and when architectural constraints participate in accounts of the system's performance is an empirical matter. Forced errors may directly account for behavioral outputs, as when the system systematically fails to accurately represent something within a domain for the lack of the necessary architectural traits.[16] This would still amount to a representational account. However, forced errors may play other less direct explanatory roles. They might be, for instance, the reason why the system stops relying on representational assets within a

[14] Cubic is not to be regarded as an empirical thesis about an actual scheme employed by the human mind. It's just a tool to present and illustrate a point.

[15] The list is from Cummins (2010b).

[16] The resulting behavior need not be bad. Misrepresentation is not necessarily misbehavior. To misrepresent something might result in apt responses or even be adaptive. False positives in representational mechanisms used to predict predator behavior are a classic example of a situation where speed and tractability matter far more than accuracy. This point will be further discussed in chapter 3.

given domain and falls back to other cognitive resources. The result is a burden to the system that can (and I think must) be interpreted as a mitigation of the explanatory role attributed to representations. Whenever there is a systematicity gap between scheme and domain, representations built with that scheme lose their explanatory leverage, for they can no longer orient the system in thinking about the possible permutations within that domain. Rather, they become assets to be disciplined, and that implies the need for either additional representational resources, or an appeal to non-representational capacities that can guide its exploitation. Thus, whenever one is trying to avoid the need for an open-ended set of additional inference-guiding information (the kind of smell that attracts FP), one has to avoid gaps between scheme and domain.

But is FP all about the structural mismatch between scheme and domain? *Prima facie*, it seems plausible to think so. The story could go like this: FP emerges in LOT-powered accounts because it employs linguistic schemes in order to handle non-linguistic domains. Spatial systematicity provides an useful example. Given a flat space with an object a in position x, and an object b in position y, a system that can cognize this state can also cognize its converse, i.e. object a in position y and object b in position x. As this kind of systematicity is not enforced by linguistic schemes, it is conceivable that a linguistic system may be able to grasp object a in position x, but not in position y. Moreover, the system can formulate all kinds of spatially implausible permutations, such as that a and b are simultaneously present on the same position at the same time, or that one of them is in two places at the same time. The underlying linguistic structure of the scheme cannot be used as a guide while coping with that domain, for it provides no architectural constraints other than the ones associated with linguistic structures. Hence, enter FP.

Tempting as it is, this story falls short of an adequate diagnostic. It implies that FP could not emerge when the system exploits linguistic schemes in order to handle linguistic domains. But FP obviously arises in this case as well. LOT's explosiveness shows itself even when accounting for the linguistic capacities of a system. Thus, while structural schemes can avoid FP by relying on their own architectural features, this is not true of linguistic schemes. There's something else about them. The upshot is that FP's real nature is yet to be fully unveiled. Unfortunately, we can't risk leaving things as they stand. If the diagnostic left some stone unturned, there can be blind spots in the suggested solution. This is the discussion to which we turn now.

2.2.3 How many schemes would you say there are?

What could be so distinctive about linguistic schemes, to the point where FP becomes their exclusive problem? A first possibility could be that the linguistic domain is open-ended. But it's easy to see that it misses the point. Non-linguistic structured domains can be equally open-ended. Cubic itself can represent an open-ended set of states. The number of 3D cubes involved (10.000) is just a kind of memory constraint. It can be further expanded by adding more 3D cubes or by enhancing the precision of the already existing ones.

A more interesting possibility regards the classic way in which we distinguish linguistic and structural schemes: while the former works by reference, the latter relies on isomorphism. In this view, FP would be the product of establishing contents through references, which implies ignoring the scheme's formal aspects, at least when attributing contents to its lexical elements. But can the distinction between reference and isomorphism really provide a clear-cut boundary between linguistic and structural schemes? The answer is no, for there are counterexamples. First, it seems like we can easily provide linguistic descriptions of structures. Is a detailed linguistic description of a map still a map? In the same vein, couldn't a linguistic "specialized theory" about a city street structure provides just the same set of constraints we find in a map? Second, we also can "picture" the logical structure of linguistic expressions. The first Wittgenstein tried to provide a whole theory of linguistic meaning relying on this possibility (Wittgenstein, 1922). Linguistic contents should be the business of language-like schemes, yet this approach would allow us to have it as the product of isomorphism. His failure has no effect on this point, for even though it cannot account for natural language, it can be made to work with some logical formalisms, and that's enough for the current needs. We have thus examples of how to forge paradigmatic structural contents through reference, as well as how to forge paradigmatic linguistic contents by the sharing of the same form.

The classical criterion seems to provide no distinction between kinds of schemes. It can't rule these counterexamples out. For instance, it provides no principled way to tell why the linguistic contents afforded by the wittgensteinian approach should be regarded as language-like. In the same vein, it preempts us from making simple claims, such as that maps are examples of structural contents or that natural language trades with linguistic contents.

Given the apparent lack of alternatives, a tempting conclusion is that there is no real boundary to be drawn. Reference and isomorphism are just two ways to achieve the same expressive power. Though we may face the empirical question of whether and how systems forge their representational assets with one or another, there's really no difference at the current level of analysis. This path leads to what I'll call a monist perspective: there is only one "real" all-encompassing, domain-neutral representational scheme. Many believe that the role of such an all-encompassing scheme can be played by classical logic or maybe some non-classical variation of it. This is no surprise, for language is usually regarded as potentially universal, given it's indefinitely extendable expressive power. In this view, any domain-related scheme, including maps, graphs and musical notation, would be nothing but an artificially constrained version of *the* representational scheme. A kind of "specialized language" that accommodates only a subset of its expressive power. The difference between specialized schemes is only superficial and has a pragmatic nature: we could describe a city's street structure linguistically, but using a special purpose "language" such as a map makes it easier for us to "read" it efficiently. Indeed, this reasoning is probably among what lead Fodor and others to postulate a LOT: given the human mind's expressive power, there has to be a general way to express anything from any domain. Otherwise, we could not account for capacities such as reasoning by analogy between distinct domains.

What this makes of the claim that structures buy us something useful to handle FP? If it's true, then the adoption of multiple schemes would not be able to buy what's needed. A specialized scheme would have no peculiar architectural properties, for they would amount to a set of constraints on how the ultimate domain-neutral scheme can be used. Architectural traits would be simply another criterion to forge frame-like structures. There could be a map-frame for map domains, a number-frame for number domains, and so on. This would bring back the issues extensively discussed in chapter 1, for now we have to tame the system's expressive power by constraining its inferential power, i.e. we have to fully specify how we can articulate our domain-targeted frames. In other words, we would be forced back to the hopeless approach of taming representational productivity by disciplining inferential productivity.

If this is on the right track, monism regarding representational schemes is fatal to the project of avoiding FP through structural schemes. Indeed, that is the reason why I've rejected structural schemes in Barth (2018). At the time I was a monist myself, and so it was just natural to think that specialized schemes were just another road to the same trap in which frames, SPACs and physiological markers had fallen. From that vantage point, whatever you think the most general representational scheme could be (first order logic, a paraconsistent variant of it or something like Cubic), there's nothing to gain from the idea of specialized "subschemes", at least FP-wise. Therefore, in order to rule out the possibility that monism is true, we'd better show what is wrong with it.

2.2.4 Representing vs. Encoding

What's the issue with monism about representational architectures? We can find a plausible answer in Haugeland's account of representational schemes (Haugeland, 1998e). In a nutshell, the problem is that it conflates two different ways in which a scheme or token is able to capture the contents of a domain: by *representing* it and by *encoding* it.[17] Now, as we have been discussing, there are essentially two distinct kinds of relationship between a representational token and its content. On the one hand, the token can refer to that content. It can do that by using compositions of symbols grounded on their contents by whatever means (functional specification, causal correlation, etc.). On the other hand, the token can be isomorphic to that content. While I'm not really comfortable calling reference-grounded meaning a kind of representation, this is not the place for a terminological quarrel.[18] Thus, at least for now, the term "representation" is being used in a sense that encompasses both informational correlation and structural matching, but this is going to change in the next chapter as we further develop the argument. What matters for now is that none of these is equivalent to an encoding.

[17] Haugeland uses the word "recording", but I'd prefer the terminology later introduced by Cummins, who argues for a similar point in Cummins (2010b).

[18] I follow here the criticism presented by Ramsey (2007).

Encoding is not a relationship between a token and its content, but a process associated to a given representational vehicle. Textual inscriptions within photographs are instances of linguistic contents structurally encoded, while a set of sentences describing a map's structure is an instance of the opposite. The famous Gödel's numbers are a way to use numeric schemes to enconde linguistic contents from other domains. Encodings are associated with recovery functions: given any scheme x, it is possible to encode it in a given scheme y and later recover x's content by using this function.

As an example, consider how softwares used for graphics design work. Inkscape is a software very similar to Corel Draw, and it works with the SVG file type. An SVG is nothing but a text file that specifies, through sentences, the size, shape, color and place of many geometric forms. Indeed, if you open the file in a simple text editor (e.g. Notepad or Vim), here's a glimpse of what you would find there (don't worry about fully understanding it):

```
<g inkscape:label="Layer 1" inkscape:groupmode="layer" id="layer1">
<rect style="fill:#ff0000;fill-opacity:1;stroke:none; stroke-width:2.446;
stroke-miterlimit:10; stroke-dasharray:none;stroke-dashoffset:0;
stroke-opacity:1" id="rect46" width="99.029762" height="76.176743"
x="29.482143" y="87.690468" /> </g>
```

When the same SVG file is opened in Inkscape, a recovery function creates a visual (i.e. structural) image on the fly, and then we can see something like figure 2. Importantly, the contents of the resulting structure are a product of its being isomorphic to a red rectangle, not due to compositional properties of the linguistic scheme in which it is encoded. The linguistic content of the SVG file is completely disjoint from the structural content there encoded.

Figure 2 – A red rectangle.

Rather than expressing the same content using a distinct architecture, encoding is akin to a file type in computers: a form in which information is stored/transmitted. Encodings enable the information to be carried or stored in distinct vehicles and in distinct forms without actually representing it. The point is not that encodings have no content, though. Encoding s can have contents (they're made using schemes), but such contents are distinct and, in some cases, irrelevant to the task at hand.

For instance, the linguistic inscriptions in the SVG file can be regarded as symbols referencing numbers or properties. But such linguistic inscriptions are not there to enable the system to make language-like inferences about numbers. They're there to enable the system to render structural contents by means of a recover function. Having disjoint contents is what makes encoding so easy when compared to the task of expressing the contents of one scheme in another. Indeed, to encode the structure of a human face using an SVG-like scheme is nothing like linguistically describing that face. We encode human faces all the time with our smartphones simply by taking pictures of them. To grasp the contents of those pictures is something else entirely. We can easily loose sight of this because it's easy for us to recognize faces in a photograph. But this is a very complicated cognitive task. Only very recently the AI community could come up with somewhat reliable facial recognition systems. And despite being a huge achievement, these systems are usually very limited in scope. In the absence of such systems, our cell phones have no idea of what are the photograph's contents: all they can handle is the way in which the pictures were encoded.

The difference between grasping encodings and grasping contents can also be found in Gödel: by shifting a digit of some Gödel number to the left, we are multiplying it by 10, but in doing so we are exploiting a systematic relation of the number domain. The result would be a reference to another number, but not necessarily another word. The exploited relationship does nothing to preserve nor to distinguish what would be a well-formed sentence in natural language or in some other logic-like formalism. In this sense, using Gödel numbers to process the sentences of some artificial language is different from using the symbol strings comprising that language. A system that can only refer to entities in the number domain (i.e. numbers) can be used to encode that artificial language, but it will be completely blind to the contents of any other domain, just like the pictures we take on our smartphones are blind to what's in the photograph.

We have thus another route through which a given specialized scheme can handle contents beyond its representational power: it may not be able to represent it, but it might nonetheless encode it. With this in mind, the reply to the monist can be summarized like this: what they regard as representational schemes with greater expressive power are actually just ordinary specialized schemes being used to encode other ordinary specialized schemes.

At this point, the monists can accept that schemes can mutually encode each other. But that is still compatible with their core claim. Even though one can encode contents, that doesn't mean one can't also represent them in a domain-neutral scheme. There seems to be plenty of examples to believe that: one can provide linguistic descriptions of human faces (as well as maps and structured domains in general), and one can also "picture" the structure underlying a given sentence, as we do in linguistics. Thus, while you can encode a red triangle in the SVG file type, you can also express its contents, as we apparently do by saying "red triangle". The rejection of the monist stance requires an additional step.

2.2.5 Against monism: the incommensurability thesis

As we have seen, to say that a scheme has a target domain is to say that it enforces certain assumptions about the domain. These assumptions are manifest as constraints in the kind of content it is able to express. In order to reject the monist stance, we can build on this and show that a distinctive set of assumptions about a domain always implies a distinctive set of representable contents. If that's the case, then the only way in which a scheme can handle contents outside its target domain is by encoding it. In what follows, I'll make this case in two steps. First, Haugeland's argument for the incommensurability of kinds of contents will be presented and discussed (Haugeland, 1998e). Rather than locating the essence of a representational kind (i.e. language-like and picture-like) in the distinction between reference and isomorphism, he grounds it on the distinctive kinds of content they can express. In a second step, I'll show how contents can individuate not only kinds of scheme, but particular schemes as well.

Haugeland's main point is that linguistic and structural schemes cannot share the same set of assumptions because linguistic scheme's contents are essentially *absolute*, while structural scheme's contents are essentially *relative*. Structural schemes are said to be essentially relative because they always assume at least two (though usually more) interdependent dimensions. We can get back to the example of a cartesian plane: we simply can't graph a point in the x axle without also expressing its position in the y axle. But it goes much further. We saw how Cubic won't allow one to express an object's determinate size without also expressing its shape and determinate location. Thus, if you try to represent the redness of a tie in Cubic, you end up expressing its shape as well. In the same vein, a picture can't represent the coffee bottle as broken without simultaneously expressing that it is smaller than the mug by the side. Underlying those examples is the fact that structural representations just *are* shapes of variations distributed alongside interdependent dimensions.

The nature of the dimensions, as well as the minimal shapes available to be distributed can vary enormously. The dimensions need not be euclidean, nor continuous as in typical planes or spaces. There can be gaps, curves, ramifications or any other kind of "deformity". As for the minimal shapes to be distributed along the dimensions, they may be simple "points", but there can also be palettes comprising flat lines, circles, squares, and so on. All of these features conspire together in order to constrain how a variation can be shaped within the available dimensions.

In stark contrast, the contents of linguistic schemes are said to be absolute because they don't assume any background comprising interdependent dimensions. They're not shapes taking place within planes, spaces or whatever n-dimensional structure. Linguistic contents capture objects and properties, but these do not constitute anything like a "conceptual space" in which a content takes place. This allows them to capture both structured (e.g. a graph measuring temperature over time) and non-structured contents (e.g. a list of temperatures) one at a time. That's why, unlike graphs of Cartesian planes, you *can* express the value of a point relative to the x axle and be silent about the value relative to the y axle. That's also why you can express that the coffee bottle is broken without simultaneously expressing its shape, size or location. In Haugeland's words:

> (...) even if for every dog there were a breed, it would make no more sense to graph breed against dog than phone number against social security number: there's no space for such a graph to take shape in. Objects have their properties one by one, absolutely, and not as part of some shape, relatively. (Haugeland, 1998e, p. 193)

In order to correctly grasp Haugeland's point, one must avoid the following misunderstanding: the absolute-relative distinction about contents is not equivalent to the atomistic-holistic distinction about schemes.

Cartesian graphs are a clear example. Given a point in a plane, its value will always be a tuple such as {x,y}. Adding a new dimension to the scheme (i.e. making it capture space rather than planes) results in a change of value for every single token. Now, given any point, its value will always be a triple such as {x,y,z}. In contrast, atomistic schemes have their basic semantic elements established independently of one another. Consider a palette comprising cartoon-like drawings of a person, a dog, and a house as its basic semantic elements, as well as a plane in which you can place compositions of these elements. Users of this scheme can thus forge contents such as two dogs and a person in front of a house. Adding a new basic element to the palette (maybe the drawing of a parrot) has no effect on any of the other basic elements. Both the cartesian plane and the cartoon-like palette are (respectively) examples of holistic and atomistic structural schemes with relative contents.

What about linguistic schemes? Can they be both holistic and atomistic? Sure. Some approaches rely on a set of basic elements whose meaning is attributed independently. Thus, whether a word "dog" means dog is independent of whether "cat" means cat or parrot. The underlying scheme is evidently atomistic. You can change the meaning of a basic element without interference on the meaning of any other. However, proponents of conceptual role semantics for representational contents (as well as proof-theoretic approaches in logic), aim at establishing the linguistic contents by holistic means (for instance, the set of possible ways in which a token can be employed is the function of its relationship with the other possible tokens). This implies that the underlying scheme is holistic. Whenever you add a new possible connection, the whole set of possible ways in which a token can be employed changes. Nevertheless, the expressed content is still absolute in both cases. Whenever a token expresses the content that the glass is broken, it can do so without assuming any background comprising interdependent dimensions, i.e. it can do so without simultaneously expressing things like its color, shape or even how broken it is).

Once we're vaccinated against this misunderstanding, we can work out some implications of the absolute/relative distinction and check whether it can buy us a ticket out of monism. First, relative contents cannot bear neither truth nor falsity.[19] Given its reliance on

[19] If we understand propositions to be the primary bearers of truth and falsity, it follows that relative contents

interdependent dimensions, the relative contents' relationship with whatever they're representing is always shaded and gradual. As such, it is subject to measure in terms of greater or smaller degrees of accuracy. We can say, for instance, that a non-blurred photograph is a more accurate depiction of its target than a blurred one. But it would be wholly inadequate to make truth-conditional assessments such as claiming that the former is true and the latter is false. Importantly, the dimensions of a scheme need not be continuous for this to hold, i.e. they're not analogical, though sometimes they're glossed that way.[20] As Haugeland remarks, the relationship between populations and generations of a biological species is intrinsically ordered (i.e. structured) and thus can be expressed with structural schemes, even though both population and generation are discrete elements (Haugeland, 1998e, p. 193).

Some might be tempted to resist this claim and insist that relative contents can be assessed in truth-conditional terms. Here's an example of how the reasoning could go: one can ground the truth-conditional assessment in sets of possible worlds, as suggested by Stalnaker (1984). The content can be said true in the possible worlds where it holds, and false otherwise. Why couldn't the current state of a structure, such as that of a red tie, be said to hold in some possible worlds and not in others? That seems to be all we need in order to make truth-conditional evaluations of the relative contents. It is definitely enough for us to answer "yes" to the question 'is the tie red?". And this is just one among the many stories that can be told. If any of such story becomes a problem for Haugeland's distinction, it will ultimately become a problem for the current diagnostic of FP.

Though this line of reasoning seems to overcome the distinction between absolute and relative contents, it is actually blind to it. Absoluteness and relativeness are not different ways to partition a single set of possible worlds. Rather, one could say that they operate with distinct sets. While absolute contents partition sets of (for instance) atomic facts that captures individual objects and properties, structural contents partition shapes distributed alongside interdependent dimensions. Thus, the issue is not that the set of possible worlds in which a picture holds would be vague in virtue of blurred boundaries. Rather, the issue is that a picture can't hold in any of them, for it works with its own non-overlapping set. Shapes distributed alongside interdependent dimensions can't capture a single possible world out of the set used by absolute contents. Consequently, there's no intersection between the contents that each representational kind is able to express. In arguing for this point, Haugeland presents the following example:

> We can, for instance, say — it is a fact — that the Earth is round. Is this not the very same "structure or feature of the world" that would be represented also by a silhouette of the Earth against a bright background? I don't think so. In the first place, the sentence identifies (the fact comprises) a particular object and a specific property of that object, whereas the silhouette identifies no object or property: its (...) content is just the overall pattern of light and

are non-propositional. But I'm trying to avoid quarrels about what propositions are. I Thank Marco Aurélio Alves for making my life easier in this regard.

[20] As an example of this point, Samuels (2010) refers to structural representations as "analogical representations", which is at best misleading.

> dark from some perspective. But further, though the sentence is entirely compatible with the Earth being transparent or just as bright as the background, the silhouette is not; and so on. (1998e, p. 194)

Haugeland's example helps us see that relative and absolute contents are incommensurable. In stark contrast with relative contents, the absolute character of linguistic contents seems to be tailor-made for assessment in terms of truth and falsity. Linguistic tokens relate to their contents in an all-or-nothing fashion. Their targets are either hit or missed, which makes them perfect to express clear-cut partitions of possible worlds, as well as truth-conditional evaluation. Therefore, Haugeland's content-based distinction implies that structural and linguistic schemes are disjoint, and that they can't be assessed with the same rule: no relative content can be deemed true or false, and no absolute content can be intrinsically measured for graded accuracy. In other words, they are incommensurable. We have thus a suitable candidate to account for what's essentially distinct between the structural and the linguistic representational kinds.

Another possible source of resistance to the incommensurability thesis is rather intuitive: it seems to be incompatible with the way we handle things in the world. An instance of this move can be found in Camp (2018): she correctly notices that we often assess maps (a structural token) in truth-conditional terms and also frequently handle book contents (which comprises large collections of linguistic sentences) in terms of overall accuracy regarding events or states of affairs. We look at the picture of a room and see *that* the drawers are open, that the clothing is all over the floor and that the room is a mess. In the same vein, we take sets of sentences as drawing an "insufficiently accurate picture" of what is the case. The details may vary, but the crucial idea is that such sets of sentences comprise theories (or concepts, or whatever) who capture distinct partitions of possible worlds. The greater the intersection between the theory's partitioned set and the set of possible permutations of the actual world, the more accurately the theory "pictures" it (i.e. shares its structure). Wouldn't these amount to counter examples to the incommensurability thesis, and consequently, to the distinctiveness of structural contents? They surely seem to indicate that we can handle structural contents in truth-conditional terms, as well as linguistic contents in graded notions of accuracy. Can we reconcile those cases with the idea of disjoint contents about representational kinds? As I see it, we don't really have to, for there is no real issue.

The heart of the matter is a confusion between the representational power of a scheme and the fully-fledged set of cognitive resources of the encompassing system. The former is a purely architectural trait. The latter comes from the system's capacity to exploit and articulate all of its cognitive resources. On the one hand, there's the expressive power of representational states and processes. On the other, there's the system's overall capacity to entertain indefinitely many attitudes. While representational content is a semantic issue, attitudes are psychological processes and states. The distinction is not always obvious, given the ubiquity of the already mentioned dictum from Fodor: to believe in p *is* to represent p. But we should not conflate these. *Pace* Fodor, to entertain a psychological state such as that of believing in p says nothing about the role of the representational contents in the making up of that state.

As an example of this point, consider again the picture of a messy room. What are its representational contents? Whatever they are, they can participate in attitude contents such as “the room is really messy” or “there’s lots of stuff out of the places where they should be”. Going further, we could maybe infer that the room is a crime scene. After all, that’s exactly the kind of mess one would expect to find after the work of someone looking for valuables in a hurry and not worried about the organization of the things left behind. But none of this is strictly manifest in the picture. Ultimately, a photograph is nothing but a bunch of light incidence structures. Those structures are the token’s bare-bones representational contents. Thus, what the encompassing system is able to exploit in such a complex structure depends on its overall capacities and its full set of available cognitive resources. Feature Detectors, environmental couplings, attentional biases are just some examples of the resources that can be employed when exploiting the picture. This is what enables it to harbor attitude contents out of structural contents: that is a room in a house, there’s house-room-typical furniture, they’re not where they’re supposed to be, and so on.

Therefore, the fact that psychological states allow for complex applications of distinct cognitive resources in general does not stand in the way of the incommensurability thesis. It implies no overlapping nor coincidence between the contents allowed by structural or language-like schemes. The upshot is that we cannot take the system’s capacity to entertain a given attitude as a clue to the underlying representational architecture being employed (if any). This faulty move from determinate psychological states to determinate semantic states is not new. The famous systematicity argument from Fodor; Pylyshyn (1988) is perhaps the most well-known example of this strategy. If we can see what’s wrong with their move, we’ll get a better grip on why we should put this kind of strategy aside once and for all.

The core of Fodor and Pylyshyn’s point is that classical language-like representational schemes have the adequate structure to account for the systematicity of though. As many have pointed out, the idea that human thought is broadly sensitive to linguistic systematicity is at least misleading.[21] What Fodor called the systematicity of thought was actually the systematicity of linguistic domains. In doing so, he took for granted that our cognitive machinery is fully capable of cognizing such domains by representing them. This is a matter of dispute, though. For instance, our capacity to detect sameness of logical form is know to be limited (Kahneman, 2011). But for the sake of the argument, let’s put it aside and agree with Fodor and Pylyshyn on this. If we accept that the systematicity we find in the human capacity to cognize linguistic domains requires a representational scheme with the very same systematicity, then we have to accept that the cognition of linguistic domains implies LOT.

But what about other domains? We do cognize linguistic domains, but we also handle (for instance) spatial domains. As Blackmon *et al.* (2001) remark, the same argument from the systematicity of the cognized domain to the systematicity of the underlying representational

[21] See, for instance, Dennett (1991).

scheme would lead Fodor to a kind of pluralism about schemes. Distinct domains would imply distinct underlying specialized architectures. This is a bullet Fodor would probably refuse to bite. Being a monist, he could insist that we need a single, domain-neutral scheme. Thus, despite the fact that linguistic systematicity is not like spatial systematicity, he could argue that we can use sentences to express (i.e. represent) spatial domains as well. That doesn't work, though, for as we have just seen, absolute and relative contents are incommensurable. All we have left, thus, is the possibility of encoding. But once we open the door to the possibility of cognizing distinct domains by encoding them, the pressure for sticking with LOT is gone. Why couldn't we adopt a distinct scheme and use it to encode linguistic domains?[22] In fact, why not get rid of representations and use only encodings all the way up, like eliminative connectionists and some non-representational frameworks suggest?

This would lead us back to a lot of familiar debates regarding distinct cognitive frameworks, but that's not the path I'm suggesting we tread. The point is simply to acknowledge that the direct move from psychological capacities to architectural traits of the underlying mechanisms is faulty. It's a bad argument for the adoption of LOT and it would be a bad argument for adopting pluralism about architectures as well. The issue about how a given domain is cognized (i.e. by representing, encoding, or both...) is empirical, and as such it cannot be settled while sitting on a couch.

We can finally get back to the core issue with the monist stance regarding representational schemes. Once we accept the distinction between representing and encoding, as well as the incommensurability that distinguishes structural and linguistic representational kinds, we can easily see what's wrong with it. The monist claims that, in dealing with distinct kinds of schemes, we're actually handling the same contents through distinct representational tools. But if what's distinctive about structural and linguistic schemes are their distinctive contents, that option is simply not in the cards. Monism has to be rejected because there's no such thing as a domain-neutral all-encompassing representational scheme that can express both absolute and relative contents. What leads one to think like that is the confusion between representing and encoding. What seemed like a bunch of counterexamples to the distinction between linguistic and structural schemes were simply examples of this confusion. The difference between encoding and representing is akin to that between taking a picture of a text and comprehending its meaning. While representation is a status of some process or state that enables it to play the role of a stand in, an encoding is a process related to the representational vehicle. Given the incommensurability of absolute and relative contents, encodings are the only way through which linguistic schemes can convey structural content and vice versa. We have finally a clear-cut distinction between linguistic and structural schemes. One that relies not on the relationship between token and content (i.e. reference and isomorphism), but rather on the nature of the content that each kind is able to express.

[22] This is exactly what we get with Tensor Products within neural models (Smolensky, 1990).

2.2.6 Individuating representational schemes through contents

Any structural scheme is distinct from another structural scheme in the sense that it has a proprietary set of primitive elements. Moreover, each of them has a distinct set of dimensions and building blocks used to graph alongside its dimensions. The dimensions themselves may also be proprietary. They can obviously differ in number (a 2D versus a 3D or n-D scheme, for instance) but that's not all: they need not be euclidean and this means that they can be curved, distorted or even gaped.

Now, when first introducing the incommensurability thesis, I've relied on the analogy with possible worlds in order to mark the distinction: absolute contents are able to partition a set of possible worlds comprising (say) atomic facts. In their turn, relative contents are only able to partition shapes distributed among a number of interdependent dimensions, i.e. a kind of structural set of possible worlds. This background of n-dimensional structures, as well as the set of basic structural elements, is distinctive of the way in which structural schemes carve out the world. That's enough to distinguish between linguistic and non-linguistic contents, but the same line of reasoning can bring us further: there's no single set of possible structural worlds to be shared by every structural scheme. To have a proprietary form is not to limit oneself to a subset of the possible shapes within what would be such a single set of structural worlds. Rather, each scheme may work with its own proprietary set of possible structural worlds in virtue of its own set of primitive elements and dimensions. That's behind Cummins' suggestion that each scheme comprises a set of *a priori* assumptions about a target domain. Its proprietary form is to be understood as sets of constraints over what can be represented and how it can be processed. The upshot is that each scheme comprises a specific set of relative contents tailor-made for a given domain.

But what exactly comprises a domain? In some examples, we've talked about 2D planes and 3D spaces as comprising distinct domains. Couldn't one also regard a plane as abstracting away from the dimension of deepness that one find in spaces? In this view, there would be a single domain (3D spaces) that is captured by two schemes, a 2D and a 3D one, but the former captures it in a less than perfect way. This is somewhat analogous to classic debates in representationalist literature regarding the indeterminacy of contents: should we describe a representational capacity as that of tracking black dots accurately or tracking flies inaccurately? Whatever one may think about this in the context of content determination - and despite the similarity in form - there's no real fact of the matter about it in the present context. As far as the explanatory role of architectural features go, it makes no difference whether we say that two schemes can capture a single domain with distinct degrees of accuracy or that those same two schemes accurately capture two distinct domains. What matters is how accurately a scheme captures a targeted domain, and the specification of the target domain is dependent on our explanatory needs. Thus, one is free to characterize a scheme X as targeting the domain of black dots accurately or as targeting the domain of flies inaccurately. Whenever scheme X is employed within the fly domain, the scheme's (in)accuracy towards that domain will be the same, and that underpins all the scheme's explanatory burden.

We can thus say that any two schemes may capture the same domain in different ways, for they enforce distinctive sets of assumptions about its form and dynamics. In other words, they can be related to each other in terms of how distinctively they capture a given domain. For instance, a scheme A can capture more fine-grained aspects that are left out by scheme B, and thus scheme B can be said more abstract than A. This means that, whenever scheme B is employed, it will result in forced errors that lead the system towards a grasp of certain features of the scheme that can be considered more abstract in comparison with the features made available by the application of scheme A. That's an example of how forced errors can be useful and informative. Science's got plenty of such examples where one learns about a complex system by employing schemes that abstracts away from the target's lower-level dynamics. Whenever a system finds out that it can articulate any two schemes through the relationship of abstractness, it may learn a lot about the target domain. This goes both ways: a system may find out that a scheme can be employed to capture features of a domain that would be otherwise out of the picture. This kind of strategy is specially useful considering the fact that cognitive resources are always limited and constrained by time and memory. One can't be expected to rebuild the structures comprising the whole world in order to represent it. Thus, in some circumstances, the possibility of trading representational accuracy for speed or tractability is valuable, and sometimes a less-than-perfect scheme might be more suited for the system's current needs than the other available ones.

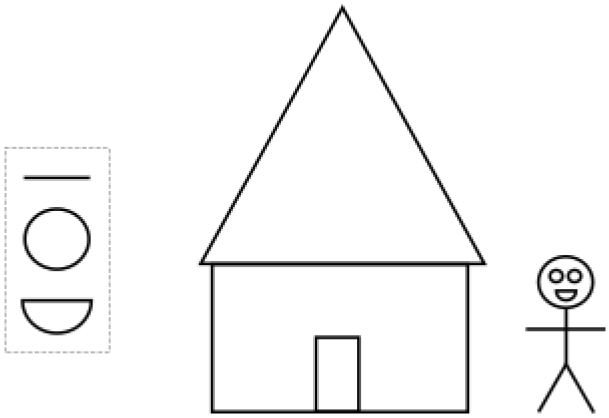

Figure 3 – Palette-C. A scheme for representing simple cartoons.

In the same vein, any two schemes can be revealed as providing relatively similar sets of representable contents regarding a domain. Consider Palette-C, an atomistic structural scheme used to represent simple cartoons (figure 3). It comprises three building blocks (a line, a circule, and a closed semi-circle) that can be arranged alongside a bidimensional plane. In figure 3, the scheme is used to forge a token representing a cartoon-house and a cartoon-person. Say there is a variation called Palette-D whose only difference is an additional building block shaped like a quarter circle. Where we to forge the exact same token with Palette-D (i.e. the quarter circle block was not employed), would it have the same content as the token built with Palette-C? Sure. This means that the contents of both schemes can overlap or coincide, probably to a considerable degree. Nonetheless, Palette-D allows for representational contents that are unavailable to Palette-C. It also makes room for a considerably larger set of possible permutations, just like adding a new word to a lexicon would do for a linguistic scheme. Thus, despite the possibility of coincidental contents, the schemes may lead the system towards very distinctive paths. For instance, while attempting to learn how to exploit its own structural traits, a system employing Palette-C rather than Palette-D would learn a distinctive set of permutations. This aspect of the learning process would be akin to that of a kid that finds a different set of combinations while exploiting distinctively shaped plastic building blocks. In this sense, the contents afforded by Palette-C and Palette-D are disjoint, for the answer to the question "what can we add to this token?" will always be different (i.e. the possible inferences will be different due to the distinct sets of basic building blocks).

Consider also what would happen if we tried to embed a token from the 2D scheme representing a square within a 3D scheme.

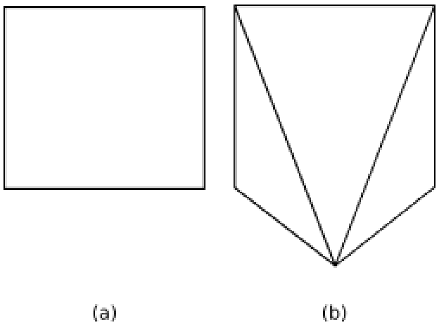

Figure 4 – Embedding 2D shapes in 3D schemes

Figure 4 (a) shows the square in a bidimensional plane. Whenever it is embedded in a 3D, the resulting shape is different, for 3D schemes enforce the need of a third dimension. Whenever you try to represent a 2D square in it, you get forced error, and the square becomes a somewhat weird 3D shape. The exact resulting shape depends on the scheme. In figure 4 (b) we can see the result of a 3D scheme that enforces the value zero. In math notation, we could say that by embedding a <1,1> square in a 3D scheme, we get a {<1,1>,0} polygon. This is a different shape distributed along a distinct number of dimensions, i.e. we have distinctive contents and, consequently, distinctive causal powers. Systems exploiting permutations of the 3D shape will reach distinct results from those exploiting the 2D shape. Thus, the contents are necessarily disjoint.

What about the opposite? Consider a cube from a 3D scheme such as the one in figure 5 (a). What would be the result of embedding it in a 2D scheme? Again, there is forced error. The cube could be rendered in different ways, such as the ones we see in figure 5 (b). These comprise distinct ways to "embed" a 3D cube into a 2D space. The most obvious possibilities for the set of axles {x, y, z} is to embed them as {x,y} or {y,z} or {x,z}. We might be inclined to say that some of these possibilities are a more accurate depiction of what we would expect

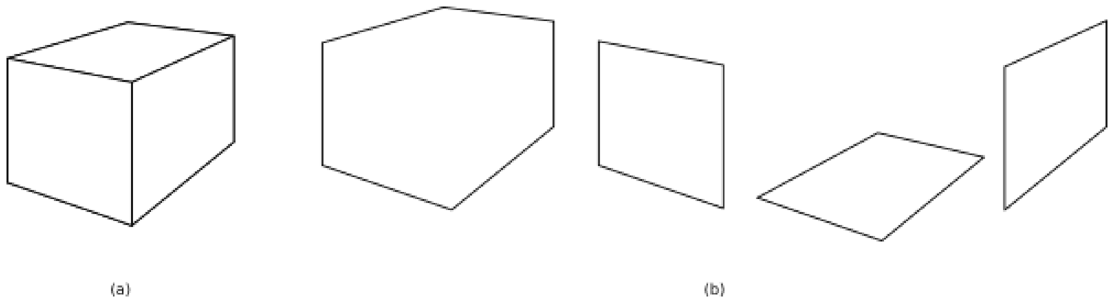

Figure 5 – Embedding 3D shapes in 2D schemes

of a 3D cube whenever embedded in a 2D space. As when one of the cube's face gets cut off and we get what would be a 2D square. But that's misleading. Another possibility that involves the whole 2D plane occupied by {x,y,z}, as we can see in the first shape of figure 5 (b). There's no principled way to claim that one embedding strategy is "better" than the other. Thus, even the apparent similarity of 2D and 3D schemes is first and foremost a product of our own exploitative capacities. As far as the underlying scheme is concerned, to carve out portions of a token in this way is no different from tokening a 3D cube and ignoring the third dimension. 2D and 3D shaped elements have quite distinct causal powers, even though we can relate them somehow ("this is how a cube should look like whenever you embed it in a 2D environment").

This is an example of how any two schemes, despite being both structural, cannot be measured with the same ruler. A scheme may do a better job regarding certain dimensions of its target domain, while some variation of it might fare not so well. For instance, while a scheme may be suitable to accurately capture size and shape, other may emphasize location and color. The upshot is that there is no exploitation-neutral criterion with which we can measure how accurately a scheme captures its domain. Unlike absolute contents, that are assessable in truth-conditional terms, there's no such thing as evaluating accuracy "from the viewpoint of nowhere". There's no sense in which we can say, for instance, that a blurred picture is less accurate than a non-colored one. It all depends on the needs of the encompassing system.

To see this idea at work, consider a 3D scheme and suppose that there is a way to measure the accuracy with which a token captures a specific dimension. Say we can express this measure with values between 1 and 10. Thus, a token that is maximally accurate would be classified as {10,10,10}, for it maxes out in all the scheme's dimensions. We certainly can compare such a token with a minimally accurate one that's classified as {1,1,1}. But what are we to say of a token A that does {10,5,10} in contrast with a token B whose accuracy is {5,10,10}? We could say that A is less accurate with respect to the second dimension and that B is more accurate than A with respect to the same dimension. What we could not claim, however, is that A is a more accurate token than B or vice versa.

Thus, in at least some cases, even two tokens of the same scheme cannot be measured with the same ruler. It doesn't mean that the contents are incomparable, though. What it does mean is that ir order to do such a comparison, the system must possess the right set of exploitation capacities. If it employs a 3D scheme to capture some domain and a 2D scheme to capture another one, it may do so without realizing that these schemes share one or more dimensions (assuming that's the case). Indeed, to realize something like that is an achievement, and not necessarily an easy one. One shall not forget that it took us a Galileo to learn that geometry can help us to think about time and speed, and a Descartes to realize that we can use equations to encode trajectories in a plane.

In the resulting picture, there's a bunch of domain-specific schemes that can bear a plurality of relationships among themselves. Given the proprietary form with which a scheme captures its target domain, there's probably no easy way to render anything like a classification system for them. Schemes can be compatible in some circumstances and incompatible in others. Such compatibility comes in degrees: one can be a more abstract version of another or a conservative variation. Depending on the system's needs, they might be either useful or comprise an obstacle. Thus, the capacity to articule them is distinct ways can make all the difference for the system. Indeed, it is possible that the system knows how to exploit any two schemes without realizing how they can be usefully articulated.

2.2.7 A final diagnostic for FP

Is this progress? Our goal, remember, was to find the source of FP and the reason why it seems to be unavoidable in linguistic schemes. One of the open questions was: assuming that FP only emerges whenever there is a gap between scheme and domain, how come that it bites even if we employ linguistic schemes in order to model linguistic domains? Now we know why. If such schemes are characterized by the kinds of contents they enable, it follows that the explosiveness of LOT is actually the explosiveness of its absolute contents. Absolute contents are never intrinsic, i.e. they bear no structural resemblance to any domain, for they have no structure. The diagnostic also helps us to understand why every classic attempt of handling FP by means of distinctive logical formalisms have failed. All of them ended up in the same place because all of them worked with absolute contents.

The point is not just that linguistic schemes do nothing to preempt permutations that are alien to non-linguistic domains (e.g. "objects always fall up"). The point is that nothing scheme-wise preempts even compositions alien to the linguistic domains themselves, such as "Pirapora book sleep John". Even though one may correctly insist that linguistic domains do have structure (morphological, syntactical and so on), the point is that such structure is not captured by linguistic schemes in any way. Consequently, the validation and exploitation of absolute contents must rely entirely on the system's capacity to enforce the adequate set of compositional rules. The system is always completely responsible for keeping them sane and useful. But the only way to do that without falling for FP and without giving up on representation's explanatory role implies an open-ended regress: we need to represent exploitation rules, which must themselves be exploited according to second order exploitation rules, and so son. This is why, as far as absolute contents are involved, FP is not just unavoidable, but also unsolvable.

We can also see more clearly the exact conditions under which structural representations (i.e. relative contents) can be of help regarding FP. Unlike linguistic schemes, structural schemes are structured. Their contents are always relative. So there is always the chance of a match between scheme and domain, i.e. structural schemes can have target domains. Whenever there's no gap between the structure of scheme and domain, architectural constraints will be enough to avoid the possibility of expressing permutations alien to the target domain. This is enough to avoid FP in that domain.

Maybe at this point, one can worry about the size of the structured domain. What if the scheme allows for an open-ended set of permutations? How can architectural constraints help the system figuring out the adequate ones? The answer is disappointing, yet illuminating: they don't. To expect otherwise is to make a confusion regarding FP and RP. Finding out the relevant permutation for the current situation is just another instance of RP. What architectural traits buy us is the possibility of organizing information in a way that can be efficiently retrieved in an open-ended set os distinct situations. That is what we couldn't do with absolute contents. Thus, while exploiting the set of possible permutations, the system need not worry about keeping up with the domain's dynamics. For instance, if the structure of a map changes, the system may need to reassess the best possible path towards its goal all over again. And that is definitely the kind of task in which the number of possible paths may require a way to be sensitive to the subset of relevant ones (i.e. it may require the system to handle RP). What the system that relies on structural representations will not have to worry is about recalculating the relationship among every single element in the map, before it even start to wonder about the best path. What's next to what, what's the relationship between theirs sizes, distances, shapes, positions, all of this will be given by architectural traits.

Having said that, it is important to be aware of what happens when the match between structural scheme and domain is less than perfect or when there is no structural match at all (such as when a scheme is used to encode content from another domain). First, there can be cases of forced error: the system is unable to handle a permutation of some domain because it lacks the appropriate set of assumptions. Second, the system may have to take over, which means it either cannot rely on its representational assets any more (which means giving up on the explanatory role of representations), or that it becomes the sole responsible for keeping those assets sane and useful. That is the same situation we're left whenever we employ absolute contents. Therefore, though relative contents seems to buy us a way out of FP, it comes at the cost of being domain-dependent. Whenever the system leaves the domain's safe shores, it is again left to go astray.

As things stand, the purchases of structural representations may seem rather thin. The explanatory role of architectural traits, as well as the possibility of representing dynamic change are deeply connected with the match between scheme and domain. But if the intrinsic way in which structural schemes capture systematic relationships works only within their target domain, how can it be of help with FP in realistic scenarios? We need the representational productivity that LOT promised us, but we also have to avoid FP and the confusion between representing and encoding. After all, humans are not single-domain creatures. If structural representations cannot avoid the potential loss of representational productivity, then the cost of its purchases will be too high. That is the crucial motivation for representational pluralism, a suggestion of how we can have non-explosive (FP-free), domain-independent productivity without relying on absolute contents.

2.3 Representational pluralism

The fundamental claim of representational pluralism is that representational productivity can be achieved through the complex articulation of a plurality of structural domain-targeted schemes. Despite first appearances, the idea is not outlandish at all. At the agential level, we already do it on a daily basis. Whenever we must draw a wire diagram, we don't use the same palette we would when drawing a flowchart. Rather, we employ a specialized palette with a very specific set of building blocks. The same goes for maps, scale models, music notation and so on.

We do have natural language (as well as formal languages) that can be employed as all-purpose domain-neutral encoding medium. But it's easy to see how inefficient they are when applied to domains for which there are specialized schemes available. Furthermore, even though we don't always realize it, we skillfully apply different sets of schemes in order to enhance their expressive power. As an example, consider graphs of cartesian planes. They cannot express the same point in two distinct positions. By drawing two points within the plane, the resulting content is not that the same point is in two different places, but that there are two distinct points. Fortunately, articulating two distinct representations built with distinct schemes enable enhanced expressive power. Thus, if one really needs to express the same point in two distinct places of the plane, some kind of identity label can be added (a symbol, a color, a drawing, etc.). Any two points with the same identity label will express the information that they're actually the same. This allows the encompassing system to communicate the idea that the very same point is occupying two places at the same time.

Representational pluralism is an application of this kind of strategy at the sub-personal level. It suggests that the expressive power of multiple schemes can be indefinitely enhanced. Distinctive mechanisms or subsystems can articulate their domain-specific representational power in ways that enhances the encompassing system's overall expressive power. This is plausible because our target phenomenon is the representational productivity of the whole system, not just that of a particular representational scheme. There's no reason to stick with the idea that the system's expressive power must be accounted for by a single scheme. Quite the contrary. The previous discussion taught us that relying on a single scheme is a dead end. Thus, rather than hoping for a single non-explosive domain-neutral scheme, one can build productivity out of the complex articulation of a plurality of less expressive schemes.

Representational pluralism comprises a bottom-up approach. Essentially, this means that as the system acquires more and more representational capacities (e.g. by learning), this process should be understood as the ability to articulate domain-specific resources in increasingly complex ways. Our goal, however, is not just that of understanding how can this be achieved. We also need to make sure that we won't get a structural version of FP as we integrate an increasingly rich set of representational resources. How can we be positive that the structural representations won't become explosive in virtue of the complex integration of multiple tokens? We've seen that linguistic systems require an unrealistic *ceteris paribus* clause. Couldn't structural representations require something along the same line in order to keep useful and consistent as they are integrated?

My answer will be no. But in order to properly justify it, we must take the long road. In what follows, I'll characterize my understanding of *representational redescription*, the learning process trough which the explanatory role played by architectural traits can be indefinitely enhanced. After properly introducing the idea, I'll argue that such enhancement implies no need for anything like a *ceteris paribus* clause, and we can have thus a kind of non-explosive representational productivity.

2.3.1 Representational redescription

Representational redescription (or simply redescription) was first presented by Karmiloff-Smith (1992). In its original sense, redescription comprises processes through which a system can learn increasingly portable contents from its own cognitive capacities. In this case, "portable" means that the content becomes available in distinct forms and can be further exploited by distinct mechanisms. It is a process through which *"(...) implicit information in the mind becomes explicit information to the mind."* (Karmiloff-Smith, 1992, p. 18). Thus, a system may be able to, say, recognize zebras in its environment, but this knowledge can be rather inflexible. As the system learns about how its own recognition capacities work, this knowledge becomes available in other, more portable forms, and these can be further exploited. Thus, the know-how involved in identifying zebras on natural environments may be redescribed as the complex articulation of mechanisms sensitive to stripes, to horse-like shapes, and so on. The resulting knowledge is more flexible and can be more easily applied to distinct environments and tasks. After all, being able to handle stripe structure can be useful for more than recognizing zebras. Here's how Karmiloff-Smith herself presented the idea:

> My claim is that a specifically human way to gain knowledge is for the mind to exploit internally the information that it has already stored (both innately and acquired), by redescribing its representations or, more precisely, by iteratively re-representing in different representational formats what its internal representations represent. (Karmiloff-Smith, 1992, p. 15)

In a nutshell, representational redescription comprises learning processes feeding on information that's already available in the cognitive machinery. The process can be recursive, which means its outputs can be subject to further redescription in an open-ended fashion. Unlike Karmiloff-Smith, I don't think there's anything particularly human in taking this kind of information as a learning *corpus*. The amount of redescription that the human apparatus does might be peculiar, as well as the level of achieved complexity. But there's no reason to narrow down its application to human animals only. Indeed, Povinelli (2012) is an example of researcher that employs representational redescription as a tool to help explain the gap between human and ape cognition. It is possible that what Karmiloff-Smith's had in mind was that redescription provides a path from structural representations towards linguistic representations. Fully-fledged natural language is definitely something that only humans managed to achieve. There was hope that we could explain human linguistic capacities by appeal to linguistic representations (i.e. representational states with absolute contents), and redescription could be a way to account for how human developmental processes allow that to happen. It should come as no surprise that I regard this as a bad thing. We already know that whenever absolute representational contents enter the mind's mechanisms in a strong non-compartmentalized fashion, FP comes with it. Though we are linguistic creatures, we cannot explain those capacities by positing absolute contents within the mind. Thus, what I regard as redescription is not a process from structural (relative) representational contents towards linguistic (absolute) representational contents. Rather, its final product is the capacity to token structural contents the system could not forge up to that point. Thus, what's distinctive to representational redescription is not its learning *corpus*, but what it produces.

An illuminating example of how the redescription process could go can be found in Ribeiro; Saverese; Figueiredo (2017). They've shown how machine learning techniques can be employed so that a system can learn about structural identity and similarity between nodes in a structure. The idea is to map the role played by individual nodes within that structure and use that role as a category. This allows the system to learn structured second-order functions. It can learn, for instance, the linkage pattern among neural sub-populations. The idea is not really new, though. It is akin to the inner workings of contemporary web search engines that attributes second-order functions to websites. As per Kleinberg (1999), some websites (i.e. nodes within the network) can work as points of reference or "authorities" to a given topic. Usually the authority is established by assessing the node's neighborhood and checking how heavily it is referenced by other sites. But what's interesting about the work of Ribeiro and colleagues is that their approach can map node functions solely in virtue of their position within the structure, i.e. by structural similarity. This means that there's no need to share the same neighborhood. It can find similar functions even in nodes very apart from one another within the structure.

Consider figure 6.[23]. Nodes u and v are structurally similar, which means that they play similar roles within the structure. Both have similar degrees (u is connected to 5 other nodes, while v is connected to 4). Both comprise a similar number of triangular connections

[23] The image is from their original paper.

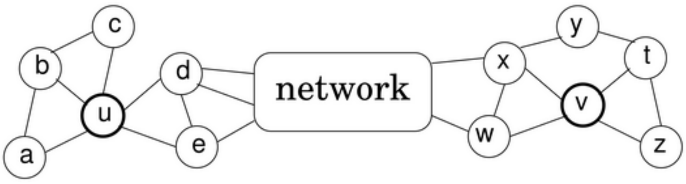

Figure 6 – Structural similarity between two nodes.

(u comprises uba, ubc and ude while v comprises vxw and vtz). Finally, both are connected to the remainder of the structure by only two nodes (v routes through x and w, and u routes through d and e). Despite the structural similarity, they're far from one another, i.e. they don't share neighbors.

This approach throws light on how a process can map second-order structures and patterns even among distinct mechanisms. It not only maps the structure's content, which is enough to make explicit a piece of structural knowledge previously left implicit. It can also map second-order roles such as that of being an "authority" within a given structure. Thus, it can learn how to map authoritative nodes within mechanisms from rather distinct domains. As it grasps this second-order structural knowledge, it can learn to structurally represent which neural subsystem typically gets along some other, and it can do that under various dimensions (temporal, spatial, etc.) and in multiple scales. The potential result of this kind of learning is structural knowledge about how the system performs under distinct situations and about the character of the cognitive resources it employs. In a nutshell, a system can learn to represent and exploit its own dynamics. As previously discussed, that's the kind of content that cannot be expressed by language-like schemes without artificially constraining its expressive power or bringing up FP. Indeed, this kind of redescriptive process is at the core of the suggestions I'll make on how to handle RP, but let's not get ahead of ourselves, for there's still a lot of hurdles on the way that require attention.

Another important distinction from Karmiloff-Smith's original work is that she employs the word "representation" in a rather loose sense. A lot of what we would regard as non-representational (such as encoded contents) would be representational in her view. We must distinguish between finding out new ways of exploiting representational resources that the system already has (i.e. increasing its inferential productivity) and becoming able to bear more representational states or processes (i.e. increasing its representational productivity). In order to understand how the mind can achieve non-explosive representational productivity, we have to know the difference between these and the specific articulations that result in enhanced expressive power. Thus, one should not understand every improvement in the overall capacity of a system to better exploit its own resources as enhancements of its representational productivity. The upshot is that, in the current framework, not every redescription is representational redescription. As we'll see, encoding, problem-embedding and analogical exploitation might allow for significant cognitive purchases, for they identify new ways in which existing resources can be exploited. They can provide enhanced cognitive capacities or increased knowledge about the world, but they can't forge enhanced expressive power. The latter is possible due to a specific kind of learning capacity, and the name "representational redescription" will refer to it.

Before further developing what's peculiar to representational redescription, let us work out a more accurate description of its learning *corpus*. What is it about the available information that enables enhanced representational productivity? Roughly, there are two kinds of sources: possible relationships between already available structural traits (i.e. schemes) and the actual application of representational tokens. The first kind was already discussed at some length while presenting the thesis that schemes can be individuated by their contents. Thus, there's no need to dig further on it. The second is particularly interesting, for it may comprise both 1) architecturally guided exploitation, such as when the system is led to a conclusion through some kind of representational forced error); and 2) cases in which the system "takes over" and articulate representational tokens without such guidance (it may see some other system articulating things in a way that never occurred to it). Let's concentrate a bit on these. After that we'll see how the same information can be subject to both representational and non-representational redescriptive processes (and why the latter falls short of what we need).

2.3.1.1 Complex token articulation as food for thought

A crucial source of information available to redescription processes comes from the articulation of distinctive cognitive resources. These articulations comprise patterns of neuronal activity from which a learning mechanism can extract part of the system's dynamics. The amount of available resources may vary enormously from system to system. Systems that make use of representations can articulate them in ways that are constrained by the underpinning architecture, which enables the possibility of learning about that architecture by exploiting distinct permutations of its representational tokens. Of course, not every use of representational resources is constrained like that. The system may also employ tokens from distinct schemes and articulate them in order to render more complex permutations. For instance, it can represent the functionally-embedded sense of a chess knight and articulate it with additional representations that capture its extra-game features, such as having a certain weight or texture. But even representational systems may feature a lot of additional cognitive resources. Sensorimotor know-how, environmental couplings and feature detectors are typical examples. All of these can be articulated with representational resources and further modulated by attentional biases (the features the system is focusing on) and affections. In particular, biological systems usually care about some of their processes, states and inputs. This flexibility brings us back to classical issues: how can a cognitive system integrate its resources in adaptive ways?

The general answer is that cognitive systems can learn. Presumably, every cognitive system has a starter-kit pool of resources, but usually these are further developed throughout the system's lifespan. They're frequently exploiting new ways of applying its resources, assessing the results and improving whenever possible. That doesn't tell us much about how they do it, though, and the amount of learning strategies may vary among systems. Recently, Dehaene (2021) presented a comprehensive account of the kinds of learning that are likely to happen in the human cognitive apparatus. There we can see instances such as learning by adjusting parameters, minimizing errors and optimizing rewards. Additionally, we're somehow able to learn by exploring distinct possibilities by restricting search spaces, even when combinatorial explosions are in place. The latter case is specially interesting, for this is the type of learning that requires the avoidance of (or a solution for) RP. We need not worry about that right now, as we'll approach the issue is chapter 4.[24] For now, what matters is to check whether and how the result of these learning processes can be further exploited by representational redescription processes.

As an example, consider the strategies through which the cognitive machinery is able to coordinate distinct portions of the brain with another one. Say a system is physically inspecting toys like a red triangle and a blue circle. It comprises mechanisms like one specialized in handling shape domains and the other specialized in handling color domains, both do that by exploiting specialized schemes. How can the system knows whether a token of |red| belongs with a token of |triangle| rather than |circle|? A possible answer is provided by Crick; Koch (2003). They suggest a framework in which what binds a |red| to a |triangle| is a shared neuronal firing rate. At the core, their point is that those mechanisms share a dynamic structure (firing rate over time), and this property can be further exploited by other mechanisms, including those capable of representational redescription. Whether their hypothesis is sufficiently plausible remains an open question, but whatever turns out to be true, it is rather unlikely that the binding of different properties would leave no traceable neuronal clues.

Another example of the flexibility available for redescription mechanisms involves more complex kinds of articulations. As some enactivists argue (Carvalho; Rolla, 2020b; Hutto; Myin, 2013), the whole system can engage in exploratory activities in the environment, and these may result in complex sets of couplings that involves different resources. This is what Piccinini (2022) called active learning. The core point is that we should avoid conceiving learning as a process that happens in isolation within mechanisms whenever they passively get some input. Enactivists and proponents of ecological psychology typically think that, by describing such couplings, one can avoid the need of anything like functionally analyzed subsystem (the kind of thing that is required is order to have representational states and processes). But active learning provides no reason to deny the possibility of such functional characterizations. Quite the opposite: whenever the world's structures are absent, it can be really useful to have some other structure as its stand-in. Just like there can be couplings among brain, body and environment, there can be couplings among a brain mechanism, the system's body and other

[24] Dehaene seems to be commited to absolute (i.e. language-like) contents in the mind, so his approach obviously depart from the view being presented here. However, none of his examples, *qua* learning strategies, are restricted or essentially committed to language-like contents nor anything like an LOT.

brain mechanisms. Both leave valuable tracks (in terms of neuronal activation patterns) about how the resources are currently exploited, and these tracks may comprise the information that further redescription processes need in order to learn what's going on.

The crucial idea is that the way we employ cognitive resources frequently yields apt informational candidates to feed learning processes like representational redescription. The application of cognitive resources is obviously subject to RP, though. The upshot is that, assuming that RP can be avoided or solved somehow, representational redescription enables the system to further increase its expressive power in a way that won't bring FP to the fore. Whenever a system learns how to do something, that may become a routine or style of handling this, and these can become the object of further representation and processing. The system can thus learn about what it knows and how it employs what it knows. If, as per redescription promises, it becomes able to exploit that knowledge through representations, it means that its expressive power can get increasingly complex in a way that does not entail RP nor FP. There's also an obvious snowball effect to the approach: the results of redescription processes can be further employed, and the relationships between them can be further exploited by redescription processes as well. This is the kind of reiteration we're after, for that's what we need in order to get open-ended expressive power, i.e. representational productivity.

However, not every learning process is a redescription process. Other kinds of learning may target information regarding relationships among structural schemes and information about how the available pool of cognitive resources is currently employed. In order to avoid rendering the idea of redescription trivial, we need to know how to spot the difference. We'll do that by quickly discussing three other approaches to this kind of information: encoding, cross domain problem-embedding, and exploitation by analogy.

2.3.1.2 Playing with food: encoding

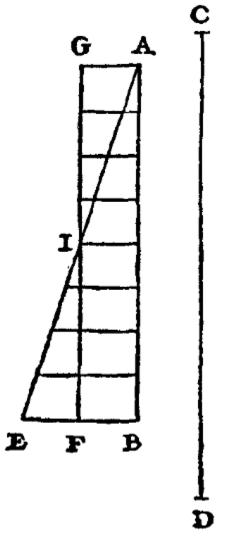

Figure 7 – Theorem 1 from Galileo

Consider now the ability to encode. In most cases, encoding a content provides no insight about it. That's expected, for as we have seen, the relationship between a content and its encodings is blind to the content's underlying structure. Nonetheless, there are cases in which we've got lucky, and we learn something from these. Here are two examples where the ability to encode ended up in significant discoveries about a domain: first, when Galileo applied geometry in order to express relationships between times and speeds. Unlike others at his time, Galileo took geometry to be more abstract. He deployed geometric tokens in order to do surrogative reasoning about non-geometric domains. His achievement, however, goes beyond the capacity to encode information. Such reasoning allowed him to establish for instance, Theorem 1.[25] By reasoning with the help of the token we see in figure 7. In other words, he found out that, at least to some extent, these domains share structure. They're isomorphic, which means times and speeds need not be linguistically encoded: they can be structurally represented as well.

A second useful example can be found in the history of imaginary numbers. As we know, $\sqrt{-1} = i$ was stipulated by *fiat* in the 16th century Italy. The number domain was arbitrarily enlarged in order to solve very specific practical issues. Nonetheless, this *prima facie* wary move appears to have hit something significant about the domain's underlying structure. The imaginary number *i* can be found, for instance, in Schrödinger's differential equation describing the wave function of quantum-mechanical systems. Explaining how this could be is a well-known and difficult problem from the philosophy of science and mathematics, so we're not going further into it. What matters for us is to notice that an apparently unwarranted move in an encoding made us open our eyes to something significant about the encoded domain.

Should we understand both examples as cases of enhancement in representational expressive power? Consider first Galileo's move. Inasmuch as a token shares structure with something else, it can be said to represent that something by isomorphism. [26] The scheme's user (Galileo) actually realized that he could represent and exploit an additional domain it could otherwise only encode. But let's not forget that Galileo already mastered (at least to some degree) the exploitation of the geometric domain. He wasn't learning how to cope with a structure that was previously unknown to him. What he realized is that he already had the

[25] *"The time in which any space is traversed by a body starting from rest and uniformly accelerated is equal to the time in which that same space would be traversed by the same body moving at a uniform speed whose value is the mean of the highest speed and the speed just before acceleration began. The time in which any space is traversed by a body starting from rest and uniformly accelerated is equal to the time in which that same space would be traversed by the same body moving at a uniform speed whose value is the mean of the highest speed and the speed just before acceleration began".* The excerpt is from Galilei (1954).

[26] Theories of representational contents that ground them in structural isomorphism (in contrast with those that ground contents in something else, such as correlational information) usually take structural correspondence to be necessary, but not sufficient. This is not a problem for my claim about the nature of Galileo's deed, for it relies solely on the fact that the isomorphism has *some* explanatory role to play in the constitution of representational contents. It need not be the whole story. The introduction of additional constraints (such that some instances of isomorphism ground representational contents, while others don't) need not concern us here. Furthermore, this topic will be properly discussed in the next chapter.

architectural means to exploit a new domain with the correspondent structure, i.e. he realized a new way to exploit representational resources that were already available for him. By analogy, that's something akin to realize a new way to use a map one already has, as if we found out that a map of city A can be used to walk around at city B because the cities share street structure. The result is not a new kind of map. Rather, it's a new successful application of an already available one. Thus, Galileo's achievement does not amount to an enrichment of his representational power. Rather, it amounts to a new way to exploit representational capacities he already had. That's definitely informative, and surely counts as an important kind of cognitive achievement. But it does not amount to an enhancement of Galileo's overall representational resources, that is, it does not amount to the acquisition of a new architectural resource.

The example of imaginary numbers enables us to extract a similar lesson, but there's an important difference. Rather than realizing that a resource we already have fits a domain other than the one we usually apply it to, what we're learning is that we can use encodings to think about things our minds can't represent. As a matter of fact, we do that all the time. We can think and talk about shapeless objects in the sense that we can make utterances that refer to these kinds of property, but we can't represent it by isomorphism. Of course, as far as imaginary numbers go, there's evidence that they do encode some underlying structure we were still not able to figure out. The fact that they don't break the systematicity of the number domain is one such evidence. Maybe we don't have the architectural means to represent whatever structure's behind it, maybe we do. The point, to reiterate, is that this is not the kind of move that can ground the enhancement of our representational expressive power. It is surely an enhancement of our cognitive capacities in general, but just like Galileo's move, it does not amount to the acquisition of a new architectural resource.

The upshot is that, though the exercise of encoding abilities allow us to learn new information about the world, it cannot be what we're looking for. Encoding is not a plausible way to increase our representational expressive power.

2.3.1.3 Playing with food: task-embedding

Task-embedding (or problem-embedding) is a kind of non-analogical cross-domain exploitation. The idea was first presented as such by Cummins (1991). The motivation for this approach is already known: given the overall incompatibility among distinct schemes - even those misleadingly similar, as we have seen when comparing 2D and 3D schemes - the permutations afforded by a scheme cannot be used as a guide to render permutations in another. In other words, inferences from a domain cannot be simply transposed to another. However, it is conceivable that a system can embed a cognitive task from one domain into another. This enables the task to be constrained by the latter. Consider a case in which a system is forging tokens of Palette-C (see figure 3). To produce such a token amounts to graphing its building blocks in a plane. But the scheme enforces no rule like "there must be something preventing the object from falling" nor anything related to the causal consequences of gravity. Consequently, the system becomes the sole responsible to enforce it, whenever it must.

But how can it learn to do so? Fortunately, in at least some cases, the system may recruit architectural help from other schemes. Tokens or building blocks from Palette-C can be embedded within a causal scheme G somewhat like we just did when discussing the relationship between 2D and 3D shapes. The essential idea is to submit the exploitation of Palette-C's features to G's constraints. Thus, G may enforce the need for causal connections among every feature, which can mean no floating entities, no changing their place without also changing place of whatever is on their surface, and so on. As an example of this point, say the system is trying to learn how to render more realistic cartoons. One of the things it would have to learn is whether moving a Palette-C-powered coffee-cup cartoon alongside the scheme's dimensions should have any consequence for the cartoon-plate beneath the cup. Assuming that G's architectural traits enforces causal relationships such as that of gravity, it won't allow permutations in which things are not grounded. Therefore, whenever the system embeds Palette-C features within G, the set of possible permutations over them is severely constrained. The task of learning how to forge more realistic cartoons partly comprises the task of finding the possible gravity-respecting permutations. In this sense, G's architectural traits may become a kind of supervisor for the system's learning, i.e. task-embedding enables a form of supervised learning.

Task-embedding is a promising and powerful learning strategy, but it falls short of representational redescription. It allows the system to learn new ways of handling things, but it does so by embedding them on what it can already express. It conflates what it knows about Palette-C's domain and G's domain, but not in a way that amount to the acquisition of new representational resources. The result may be a more efficient way to exploit Palette-C in certain circumstances, but this efficiency is achieved by constraining the exploitation of its representational power, not by expanding it.

2.3.1.4 Still playing with food: analogical reasoning

A rather similar conclusion comes from reflecting about the ability to coordinate distinct models or tokens through analogical reasoning. Thus, a quick remark will suffice. As I understand such skills, they're another kind of cross-domain reasoning that allows us to grasp entities or states from a domain by using tools whose target is another domain. If we take this in a sufficiently broad sense, then it is no different from what Galileo did: he regarded the relationships between speeds and times as akin to those of a geometric domain.

Of course, the upshot is rarely that of Galileo. Most of the time, analogies are shallow and capture only a very specific aspect or dimension from distinct domains. We can say, for instance, that electrons orbit a nucleus much as planets orbit a sun. Modern quantum mechanics has shown that this view is problematic and ultimately false, but the point remains: early attempts to model atoms were inspired by the solar system and its dense nucleus (the sun). This kind of move may result in new cognitive knowledge about the world, but that's not achieved through the enhancement of system's expressive power. Rather, the system is learning new ways to apply representational resources it already has.

2.4 Explosion-free representational productivity

Let's quickly take stock. Human world contexts comprises situations where complex coordination of cognitive resources is needed, particularly representational ones. Therefore, the remaining fundamental question is whether we're allowed to keep structural scheme's purchases as we walk towards increasingly complex applications of the system's resources. Commonsense requires the capacity to recognize and handle individuated objects as such, even though we think about them through multiple tokens that capture distinctive aspects of them. In a chess game, a piece on the board is both a knight and a small statue made of wood or stone. Being able to grasp such an object in distinct ways at distinct times requires some kind of integration among the cognitive resources employed. Otherwise, we would not be able to plan a chess move with a knight and take into consideration the weight of the piece while executing the move. Furthermore, while discussing situation holism, we saw that understanding human-world situations involves articulating and applying distinct models of the situation. Such models might be the product of assembling different tokens from distinct schemes, and if the set of possible situations is open-ended, then there has to be a productive way of doing so. The complexity involved can escalate very quickly, for thanks to situation holism, we can grasp the same situation in more than just one way, as we do when playing pretend. Representational pluralism makes this the burden of representational redescription processes.

In the previous sections, two of the core informational sources for redescriptive learning processes were presented: the system can exploit relationships between structural schemes as well as the way cognitive assets are currently exploited. This information can become the learning *corpus* of both representational and non-representational redescriptive processes. Non-representational redescription comprises processes such as encoding, task-embedding and exploitation by analogy. The output of such processes is enhanced inferential productivity, i.e. the system can find out new ways to articulate its current set of cognitive resources. Roughly, it results in new cognitive knowledge. For instance, if the system could already process tokens of points in a plane, and could already think about color as well, now it might be able to reason in terms of colorful points in a plane. That does not amount to a representational kind of redescription, though. In order to qualify as representational, the outputs of the redescription processes must retain representation's explanatory purchases. But such representational purchases are a feature of their underlying schemes. More specifically, they're a feature of how accurately a scheme captures a domain. Therefore, the output of representational redescription must be a new scheme, i.e. a new set of architectural assumptions about some target domain.

It follows that even unconstrained exploitation of cognitive resources (representational or not), that proved successful or adaptive enough to be stored, can be further redescribed as a new representational scheme. Rather than merely reenacting some already stored knowledge, the system can now exploit architecturally constrained permutations of that knowledge. In other words, it can forge representation-guided extrapolations of its previous exploitations. Thus, the redescribed scheme comprises a more *portable* rendering of whatever the system learned about the world.[27] And this rendering is stored as a set of architectural constraints, which means it can guide further reasoning. In a nutshell, representational redescription opens up the possibility of forging architectural constraints from the way tokens of different schemes are applied, exploited or organized. The system's representational capacities are effectively enhanced. We have thus a path towards open-ended representational productivity.

That being so, we should not conceive of representational redescription as a learning strategy or a kind of learning. It is rather a peculiar goal: increased portability. This preserves the spirit of Karmiloff-Smith's insight, even though it rejects the idea that increasing portability is a step towards linguistic representational schemes. Rather, it goes towards more portable structures, as previously exemplified through the work of Ribeiro; Saverese; Figueiredo (2017). But why would a system pursuit this goal? Its real value regarding RP won't be fully unpacked until chapter 4, but some hints are already in order: increased portability is a way to do more with less. It is a way to trade complex problems for more frugal and simpler ones, so that even a limited processing power can handle it. Whenever a system applies a given structural scheme in order to handle something, the scheme "dictates" how the system takes that something to be. The system can then employ exploitation strategies that would not be available otherwise.[28]

With this at hand, we can finally get back to the issue of the potential need for a *ceteris paribus* clause in order to avoid a representational explosion as the system gets more and more complex. How can we formulate such a clause in a pluralist system? The good news is that we don't have to. Representational pluralism can accommodate an unbounded enhancement of the system's representational resources without losing track of what keeps them under control. Say a given scheme X results from representational redescription of the concomitant previous exploitation of schemes A, B and C. The fact that X derives from those schemes doesn't mean that the system loses the ability to exploit them. Representational redescription is like baking a cake in the sense that it produces something new from the complex articulations of its ingredients, but it's unlike it in the sense that its ingredients are not consumed. They're still there to be exploited just like they were before. Thus, redescription does not entail loosing sight of the architectural constraints used in the "ingredients". The tendency to think otherwise is perhaps attributable to the tendency of understanding biological mechanisms through the eyes

[27] The importance of increasing portability is also addressed by Ismael (2007). Though she's worried about more specific issues and develops her reasoning using different tools, I believe that much of what is advocated here is compatible with her work.

[28] This will be further discussed at the beginning of the next chapter.

of engineering rather than wearing evolutionary glasses. Evolutionary accounts of biological capacities are full of examples in which distinct features add up to constitute a new capacity without concealing or getting rid of the previously available ones.

Therefore, no *ceteris paribus* clause is needed, for what we want is given by the *absence* of certain possibilities due to the system's assumptions about its target domain. In other words, if there's anything remotely similar to such a clause, it is given by what's left out by the set of scheme's assumptions about their respective target domains. What the system regards as ordinary while handling its representational resources is a function of its set of architectural constraints. Which of such constraints will be effectively taken into account at any given cognitive task is a function of the scheme's being exploited. There is no need for additional data structures nor to render any assumption explicit in any way, for they're already embedded in the scheme's make up. That's why the system can readily acknowledge that a ball will fall *unless* it is glued to the roof or *unless* it is full of lighter-than-air gas, but it only does it if and when this possibility is brought about somehow (e.g. by someone else). In other words, the system may have the capacity forge representational contents of balls falling up or being glued to the roof, but it need not repeatedly check whether that's the case. Unless something brings these possibilities up, it will just stick to the domain's assumptions architecturally stored, cope with things on that grounds and preclude alien permutations.

To see this at work, suppose there's a scheme that captures some set of causal relationships with reasonable accuracy. That could work as a partial explanation of why we don't get lost in simple situations within the scheme's target domain. Given a floor and a ball in the air, we can promptly say towards which direction the ball goes once released: down. We can do that because such a permutation can be architecturally enforced by the scheme employed in representing the causal domain. Of course, we can also understand the exceptions: what if the ball is tied to a roof or is full of some lighter-than-air gas? But these considerations come from outside the core dynamics of that domain. Thus, architectural traits amounts to non-arbitrary principles that are useful when applying a token and thinking about a situation. Unless someone or something brings up such alien possibilities (relative to the domain in question), they're simply outside our horizon and regarded as non-ordinary.

Before moving on, it's elucidating to emphasize how distinct this is from the strategy of taming inferential capacities by using frame-like structures. Using multiple schemes to build-up increasingly powerful representational capacities allow us to preempt the system from going astray by relying on its multiple architectures. The representational constraints are, in this sense, a given of the way the representational capacities are specified. In contrast, frame-like structures try to artificially impose representational constraints by establishing inferential limits in how the system is able to exploit its unbounded representational resources. Consequently, it must establish "a priori" everything the system should or shouldn't infer in every possible context. By relying on architectural traits, rather than constraining the representational capacities top-down, we-re building them bottom-up.

If we're on right track, the results we get so far allows us to find the right balance between the freedom to go beyond what's present and manifest and the world's stability (as when we make predictions such as "what will happen if..."). The system can wander without loosing track of the world's ordinariness. We are finally able to get representational productivity while avoiding FP, and free to face the challenge posed by RP without giving up on representational resources. Therefore, representational pluralism can provide adequate grounds for a sleeping dog approach to FP. The set of assumptions from the system's schemes comprise principles of ordinariness that enable a kind of default reasoning. Cognitive knowledge representationally redescribed and stored as schemes requires no additional data structures to guide their exploitation. Like any other scheme, they can architecturally enforce its set of assumptions and guide the system in its reasoning by leaving out all the permutations that are alien to that domain (things like balls falling up). We can thus dodge issues that affect absolute (i.e. linguistic) contents, such as that of finding out what can be ordinarily ignored and what cannot.

The resulting picture is one in which a system can reiteratedly redescribe domain-specific cognitive knowledge. It can not only find new ways to grasp things within the domains it already targets, but also articulate features from multiple domains and forge new schemes out of those articulations. All of this can be done without giving up on the benefits of relying on representations. We can thus see how a system can increase its representational expressive power without giving up on the domain-specific benefits. Given representational redescription, even a system with a rather small set of representational resources (i.e. schemes) in its starter kit is able to reach an open-ended expressive power.

2.5 Getting ready for what comes next

Now that we have representational productivity properly established without the risk of representational explosiveness, we've made a fundamental step towards handling RP without leaving representationalism. But before we get back to RP, there's still a fundamental question that must be dealt with: what's the psychological plausibility of representational pluralism? As Waskan (2003) would put it: how do we go from the idea of scale models as metaphors for the mind's inner workings and reach a plausible mechanistic implementation of structural representations? This question can be unpacked in lots of complex issues. A complete answer to all of them requires a lot of empirical work, to be sure, but if we want the proposed framework to be taken seriously, we need to show the potential of its core tenets.

For instance, we're the kind of creature that can bear propositional attitudes about the world, but if representational pluralism throws language-like representational contents away, then we must explain how a bunch of non-propositional representational contents can add up and constitute propositional attitude contents. In doing so, we want to stick to everything that representational explanatory frameworks can buy us, and we don't want to become vulnerable to anti-representationalist arguments that would render trivial or useless everything that was discussed in this chapter. Furthermore, we have to face the need for a plausible semantics for all of those representational schemes: in virtue of what do they get their contents? How exactly a given scheme gets to say that it has a certain domain as a target? We need a theory of representational content that is, at the same time, able to answer that and preclude the typical threats posed against representational approaches. As we'll see, representational pluralism can be a very picky *desiderata*, and the vast majority of the theories on the cards won't do. Having said that, I think that the answer is already available in cognitive science's literature. All we have to do is show how to articulate them with what we already discussed.

2.6 Paragraph-by-paragraph summary

Each entry below summarizes a paragraph of the main text.

The challenge of representational productivity

Productivity of thought is usually regarded as one of the core cognitive psychology's *explananda.*

This capacity is usually accounted for by employing productive representational schemes such as LOT and relying on their compositionality.

However, LOT brings with it a lot of well-known challenges that might be insurmountable.

The one I regard as fatal is that LOT, despite providing representational productivity, preempts inferential productivity.

That's because LOT is explosive: it requires an open-ended set of additional cognitive resources in order to prevent the system from going astray.

Consequently, LOT precludes the representationalist from exploiting the full potential of representational assets.

The frame problem is a manifestation of LOT's explosiveness, and its history of failed solutions will help us see that the issue is probably unsolvable.

For the want of a frame

The frame problem was born in the AI community, and it was interpreted in many ways.

In its first guise, it was about modelling the effects and non-effects of actions or events in a logical formalism called Situation Calculus.

Situation calculus required all of them (effects and non-effects) to be determined at design time, but as they can change depending on the actual circumstances, the designer has to fully determine them for every possible circumstance *a priori.*

The need to avoid this prohibitively long list of axioms in specifying actions or events was dubbed the frame problem.

Patrick Hayes took the problem to be essentially that of modeling a principle of *ordinariness*, i.e. a princile that could stand as a *ceteris paribus* clause in every inferential step.

Despite appearances, even in its original formulation, FP is not just a problem for designers of intelligent agents, for even if we could come up with all the required axioms, the result could be very inneficient.

Inference becomes cognitively costly, for frame axioms don't allow efficient and flexible information retrieval under an open-ended set of possible situations.

This has an obviously unrealistic consequence for learning in such systems: in the absence of a *ceteris paribus* clause, learning about X necessarily amounts to learning about X's relations to everything else in the (agent's) world.

Broadly speaking, the literature presents us two kinds of strategies to handle FP: 1) the "cheap test" approach tries to reduce the number of hypothesis considered by classifying and compartmentalizing the information (this suffers from the same issues plaguing the frame systems presented in chapter 1 and won't be further discussed); and 2) default reasoning (sometimes called the sleeping dog approach).

Sleeping dog approaches appeal to default reasoning, but it only reformulates FP as the problem of determining the boundaries of what's default and what's not: how can the system know whether it has good-enough motivation to consider a non-default inferential path?

The sleeping dog approach triggered the development of non-monotonic logical formalisms, which has value on their own, but provided no substantial progress FP-wise.

Now that we're acquainted with FP's early manifestation, we can discuss its nature and relationship with RP.

What is the frame problem about?

First, we must understand why FP and RP are not the same.

Dennett was probably the first to interpret FP as an instance of RP, for he thinks that a theory of ordinariness and a theory of relevance can play the same role.

But his reading conflates issues about representational productivity and inferential productivity: it tries to tame the former by constraining the latter, and the story of failed attempts to handle FP through cheap test and sleeping dog shows that this leads to a dead end.

Furthermore, issues regarding ordinariness cannot be reduced to problems about relevance: an ordinary non-effect can be circumstantially relevant.

For the same reason, an issue about relevance cannot be reduced to a problem about ordinariness, which means ordinariness and relevance have distinct *explananda.*

Some reject Dennett's account, but only due to an excessively narrow conception of FP as the practical problem of reducing the number of frame axioms to be written down by the designer while modelling the world.

They think non-monotonic logic is all we need for achieve that, but this view is silent about what matters most: the need to handle non-saturable contexts.

My view is in the middle: solving FP is indeed about being able to model world changes efficiently, but this has to be done in a way that does not stand in the way of handling RP in representational frameworks.

The history of FP is the history of trying to model changes using LOT, which requires handling LOT's explosiveness of representational productivity, which can only be done by constraining inferential productivity, which seems impossible for non-saturable contexts.

But perhaps we can pursue representational productivity without LOT, and I'll argue that we can buy that if we rely on an alernative account that employs structural representations dubbed representational pluralism.

Structural representations and the frame problem

For now I'll abstract away from psychological, implementational and foundational issues regarding representations, which means we'll not worry about what grounds semantics, whether they're mechanistically plausible, etc.

This is a purely methodological move, for we need a clear boundary between representational and inferential capacities.

In order to avoid *ad hoc* stipulations such as "gastric juice represents the prey's scent", we'll take representations to be tokens that function as stand-ins in virtue of an encompassing scheme.

We can temporarily rely on the classical way of distinguish structural and linguistic representational schemes: the nature of the relationship between tokens and contents.

In linguistic schemes, the representation relationship boils down to that of reference: to say that a represents b means that a refers to b.

In structural schemes, the representation relationship is grounded in isomorphism: to say that a represents b means that a is isomorphic to b.

With this distinction in mind, it's time to dig further on the purchases allowed by structural representations.

What structural representations buy us

Structural representations allow for non-explosive representational assets, which enables us to avoid FP when exploiting them.

To see how, consider Erik Janlert's requirement: conservative changes in the world must involve only conservative additions to the representation.

Linguistic representations fail to satisfy it because even conservative changes yield an explosive need for additional representational assets.

This is not the case with structural representations, as we can see in the example of map-like depictions of geometrical relations among cities.

When a new sentence is added, in order to figure out what needs updating, linguistic systems require explicit tests even of what's implicit in every entry, which results in the need for an explosive amount of additional information, for without it the system can't keep the set of entries consistent and useful, i.e. the very thing that should be guiding the system needs guidance from the system.

In constrat, structural representations collapse the linguistic implicit/explicit distinction: their contents can guide cognition as if they were explicit, but their updates happen "automatically", as if they were implicit.

Thus, adding new features to a structural representation don't require making additional sets of rules explicit in order to keep the representational token capable of guiding cognitive tasks.

Representing change

Structural representations enable an additional purchase that's unavailable for linguistic contents.

With language-like schemes, the world's dynamics is typically modeled as sets of sentences depicting a clear-cut snapshot of a situation, and rules (laws) governing the transition from one snapshot to the next.

In contrast, structural schemes enable the change itself to be pictured, i.e. rather than calculating changes, the user can represent the dynamics.

In doing so, representations can get their explanatory burden back: they can guide the system while thinking about the dynamics of the target domain.

We can thus overcome the need to carve out the world as sets of time-instants and think of it in terms of trajectories, intervals, stories, etc.

This strategy is not available to language-like schemes because the carving out of those intervals also rely on architectural features of structural schemes, such as the collapsing of the linguistic implicit/explicit distinction.

Intrinsic contents in structural representations

This raises crucial questions: how can structural representations do that? Are these features possessed by every structural scheme or just some of them?

The answer to the first question is that structural representations intrinsically mirror relationships among the represented features.

But a scheme can only do that when it shares structure with the domain being cognized, otherwise the responsability for keeping the representation consistent goes back to the system, and the representation's explanatory role is threatened (just like language-like schemes).

Thus, the intrinsicness that characterizes structural representations relies on the structural resemblance between the underlying representational scheme and a given domain, which means that schemes should be considered domain-specific.

The upshot is that architectural traits can play a larger explanatory role than it's usually acknowledged in representational accounts of cognitive capacities, for much of representation's usefulness comes from the domain-targeted character of its underlying scheme.

The explanatory role of architectural constraints

Architectural traits of representational schemes can constrain both how the content can be processed and what can be represented.

Buth such purchases happen only with schemes that accurately capture their target domain, and only when reasoning within that very domain.

Architectural limits to what can be represented are called forced errors; these are different from other kinds of representational error because they're independent of the system's capacity to exploit the scheme.

Cummins suggests that these content-constraining properties are what determine the scheme's target domain, for they comprise the scheme's set of a priori assumptions about the domain.

He illustrates this point with an 3D artificial structural scheme called *Cubic*.

Cubic's architectural traits can ground spatial assumptions such as "evey object has a determinate size and shape" and "no two objects can occupy the same place at the same time" even though these are nowhere explict in the scheme's specification.

Such set of assumptions comprises the scheme's essence.

So far, it seems like whenever one is trying to avoid FP in representational systems, one has to avoid gaps between scheme and domain, for such gaps bring back with themselves the need for additional explicit reasoning rules.

This suggests that FP's ultimate source in representational systems is the mismatch between scheme and domain, for only in these cases the need for additional reasoning rules comes to the fore.

But that can't be right, otherwise using linguistic schemes in linguistic domains would not trigger FP, which means we need to look for a better diagnostic of FP elsewhere.

How many schemes would you say there are?

We know that FP won't affect structural schemes employed within their target domains, but what could be so distinctive about linguistic schemes and their respective linguistic domain, to the point where FP becomes their exclusive problem?

The heart of the matter is in the classic conception of what distinguishes linguistic and structural schemes (which we have been assuming so far): the nature of the relationship between token and content: reference and isomorphism.

This conception provides no clear-cut boundary for there can be structural accounts of linguistic contents (remember Wittgenstein's project) and linguistic accounts of structures (a linguistic description of a map).

A tempting conclusion would be that there's no real boundary to be drawn: they're just two ways to achieve the same expressive power, a position we'll call *monism*, which is common among those who think that structural schemes are nothing but "specialized formal languages".

But if monism is right, then structural schemes can't buy us a way out of FP, for architectural features would boil-down to rules regulating the exploitation of linguistically described structures, which is no different from the frame-like structures discussed in chapter 1.

Thus, in order to justify the adoption of a pluralist perspective, we must first understand why monism must be rejected.

Representing vs. Encoding

The problem with monism is that it conflates two distinctive ways in which a scheme can account for contents: by representing it and by encoding it.

Encoding is not a relationship between a token and its content, but a process associated to the representational vehicle.

We can find an example in graphics software who store the drawings in the form of complex sets of sentences that describe what the system must do in order to render the picture (a "pragmatic" approach to pictorial contents, if you will).

Thus, rather than a way to express the same contents using a distinct architecture, an encoding is akin to a file type in computers: a form in which information is stored or transmitted.

Exploiting encodings and exploiting contents is different, just like exploiting a linguistic sentence is distinct from exploiting its correspondent Gödel number.

The pluralist reply to monism is that what they regard as a domain-neutral scheme with larger expressive power is actually an ordinary specialized scheme encoding contents from other equally specialized scheme.

But the monist can still try to resist by insisting that the possibility of encoding a content does not preempt the possibility of also representing it, which means we need an additional step before fully rejecting it.

Against monism: the incommensurability thesis

We can dodge the monist's resistance by showing that a distinctive set of asssumptions about a domain *always* implies a distinctive set of representable contents.

In this connection, Haugeland argues that the content of language-like schemes is essentially absolute, while structural scheme's contents are essentially relative.

Structural representations are said relative because they're always shapes of variations distributed alongside two or more interdependent dimensions.

The nature of the dimensions, as well as the minimal shapes available to be distributed can vary enormously.

In contrast, the primitive elements of language-like schemes don't assume any set of interdependent dimensions and can be aggregated one by one.

Before going further, a quick but important remark to prevent confusion: the distinction between absolute and relative contents is not equivalent to the distinction between atomistic and holistic schemes.

A holistic scheme is one in which the contents of any token is always a product of its whole set of primitives (add a third dimension to a cartesian plane, and every token x,y is now necessarily x,y,z), while atomistic schemes have their basic semantic elements established independently of one another.

This goes for linguistic schemes as well: they can be obviously atomistic ("dog" means dog, independently of what "cat" means), but conceptual role semantics (of proof-theoretic approaches in logic) can establish absolute contents by holistic means.

Now that we know the difference between holistic/atomistic schemes and relative/absolute contents, we can appreciate that, while absolute contents partition sets of (for instance) truth-assessable atomic facts capturing individual objects and properties, structural contents can only partition shapes distributed alongside interdependent dimensions.

But couldn't relative contents be truth-assessable as well? Stalknaker, for instance, suggests that a picture can hold in certain subsets of possible worlds, thus grounding truth-conditional evaluations of it.

But that is not so: absoluteness and relativeness are not different ways to partition a single set of possible worlds, for each kind of content operates with its own distinct set.

Therefore, absolute and relative contents are incommensurable, and the distinctiveness of each kind of content can account for what's essentially distinct between structural and linguistic representational schemes.

Some may try to resist the incommensurability thesis by presenting possible counterexamples: sets of sentences may present an "innacurate picture" of something, and pictures may ground truth-conditional assessments ("the drawer is open").

But examples like these rely on the conflation of the expressive power of a particular scheme and the fully-fledged set of cognitive resources available to a system.

Therefore, the fact that fully-fledged psychological states allow for complex applications of distinct representational resources in general does not stand in the way of the in comensurability thesis.

Another instance of a similar faulty move from psychological states to semantic states can be found in Fodor and Pylyshyn's systematicity argument.

They assumed that the systematicity of thought implies an isomorphic systematicity for the vehicle of thought, but that's not necessarily so, for the systematicity of a domain can be cognized not only by representing it, but by encoding it as well.

Furthermore, even if their argument were sound, it would imply pluralism rather than LOT, for the systematicity of each domain would imply its own specialized scheme.

These examples remind us that the direct move from psychological capacities to the architectural traits of the underlying mechanisms is faulty.

The distinction between absolute and relational contents is able to resist complaints like these and stand as a plausible account of what distinguishes linguistic and structural contents, as well as a plausible reply to the monist: there's no such thing as an all-encompassing representational scheme that can express both absolute and relative contents.

Individuating representational schemes through contents

A natural development of the incommensurability thesis is that every structural scheme has a proprietary form. Any structural scheme is distinct from another structural scheme in the sense that it has a proprietary set of primitive elements.

Each scheme work with its own proprietary set of "possible structural worlds" in virtue of its own set of primitive elements and dimensions.

This means that, inasmuch as representations are involved, how (in)accurately a scheme captures a domain will always have some explanatory role to play.

A less-than-perfect scheme amounts to a set of assumptions (about its target domain) and these assumptions may be what's behind the tractability of some structure that is out there.

This also means that even though some overlapping between similar schemes is possible, any two schemes will always be overall incomparable in accuracy.

For instance, embedding a 2D square in a 3D scheme results in disjoint contents.

The same happens if we embed a 3D cube within a 2D scheme.

Therefore, from the viewpoint of nowhere, there's no way to tell whether a given content in a 3D scheme can play the same explanatory role that a token of a 2D scheme would, for that depends on why and how the content is exploited by the system.

The same reasoning can be extended from structural schemes to particular representational tokens: given any two pictures of the same scheme, sometimes it may be impossible to say which one is more accurate in general, for one can be more accurate in one dimension and the other can be more accurate in another dimension.

There's no system-neutral sense in which we can say, for instance, that a blurred picture is less accurate than a non-colored one.

In the resulting picture we have a content-based account of representational pluralism: there's a plurality of domain-specific schemes that can, to varied degrees, mutually encode each other.

A final diagnostic for FP

Now we can finally see where FP comes from: the explosiveness of LOT is actually the explosiveness of its absolute contents, for they have no structure, and thus they can't bear intrinsic structural resemblance to any domain.

Consequently, the validation and exploitation of absolute contents must always rely entirely on the system's capacity to enforce the adequate set of compositional rules, thus rendering FP not only unavoidable, but also unsolvable, for the asset that should be guiding the system will always need guiding from the system.

In startk contrast, structural schemes can architecturallyh enforce its own "usage rules", precluding FP within their target domain.

This is a good time to remember tha FP is not RP: architectural traits won't help in choosing the relevant permutations, i.e. it won't help to choose among the possible paths in a map, but it will avoid the need to recalculate the relation among every single element of the map at every inferential step.

Therefore, though relative contents seems to buy us a way out of FP, it comes at the cost of being domain-dependent, for whenever the system leaves the domain's safe shores, it is again left to go astray.

The problem is evident: humans are not single-domain creatures, so we'd better understand how we can achieve representtional productivity out of a plurality of domain-specific structural schemes.

Representational pluralism

Representational pluralism claims that representational productivity is achieved through the complex articulation of specialized schemes.

At the agential level, it is commonplace to skillfully apply specialized schemes in order to enchance their expressive power, e.g. we can label points in a cartesian plane ir order to express the same point in different places (which is impossible without labels).

The strategy can be applied at the sub-personal level as well.

As the system learns new ways to articulate tokens from distinct schemes, it gets enhanced expressive power (i.e. it moves towards representational productivity), but how far can this approach go before we need something akin to the *ceteris paribus* clause required for linguistic systems?

I'll claim we can go all the way up indefinitely, but in order to justify that, I'll need first to characterize my understanding of a very specific learning process called representational redescription.

Representational redescription

Karmiloff Smith first presented the idea as a process through which implicit information in the mind becomes explicit information that can be further exploited by the mind.

In her view (expressed in my terms), this is how we get to the point where we exploit linguistic absolute contents within the mind, but it should come as no surprise that I reject the idea of representational redescription as a move from relative to absolute contents.

Outputs of representational redescription processes always comprise more relative contents, but they may be represented through a distinct scheme that involves patterns and structures that the system was (up to that point) insentitive to.

The potential result of this kind of learning is structural knowledge about how the system performs under distinct situations and about the character of the cognitive resources it employs, i.e. a system can learn to represent and exploit its own dynamics.

But in order to fulfill our needs, we must carefully distinguish between finding out new ways of exploiting representational resources that the system already has (i.e. increasing its inferential productivity) and becoming able to bear more representational states or processes (i.e. increasing its representational productivity).

Before further developing what's peculiar to representational redescription, let us work out a more accurate description of its learning *corpus*: what is it about the available information that enables enhanced representational productivity?

Complex token articulation as food for thought

Complex token articulations comprise patterns of cognitive activity from which a learning mechanism can extract part of the system's dynamics.

The availability of this data is a bi-product of the system's capacity to integrate its resources in adaptive ways (an issue we'll discuss in chapter 4).

As an example, consider the strategies through which the cognitive machinery is able to coordinate distinct portions of the brain with another one, e.g. when it associates a token of |red| with a token of |circle| in order to render a |red circle|.

Of course much more complex processes also yield useful information, such as when the agent engages in active learning or exploration of its surroundings.

The general idea is that the way we employ cognitive resources frequently yields apt informational candidates to feed learning processes like representational redescription.

However, not every learning process operating over the system's own dynamics is a representational redescription process, and in order to avoid rendering the idea of redescription trivial, we need to know how to spot the difference.

Playing with food: encoding

First we need to distinguish redescription and encoding, for though encoding can increase our cognitive knowledge, it provides no additional representational expressive power, as we can see in the case of Galileo's theorem 1.

Another example of the same point can be found in the history of imaginary numbers.

Both are examples of cognitive achievements that do not result in enhanced representational expressive power, but rather in realizing that the system can use a representational scheme already available in a new way.

Learn to encode a domain is surely an enhancement of our cognitive capacities in general, but it amounts to learning that an already available representational asset can be employed in a new way.

Thus, though encoding allow us to learn new information about the world, it is distinct from representational redescription.

Playing with food: task-embedding

Task-embedding (or problem-embedding) is a kind of non-analogical cross-domain exploitation.

The basic idea is that a system can subject a representational token from a scheme C to the architectural constraints of a scheme G.

Thus, G's architectural traits may become a kind of supervisor for the system's learning about C's target-domain, i.e. task-embedding enables a form of supervised learning.

But again, though this allows the system to learn new ways of exploiting already existing representational resources, it does not amount to increasing the system's representational capacities.

Still playing with food: analogical reasoning

A similar upshot comes from metaphors and analogies: these are another kind of cross-domain reasoning.

They only allow for new ways to exploit expressive (representational) resources the system already has.

Explosion-free representational productivity

The final product of redescription processes are new representational schemes with new architectural traits.

Enconding, task-embedding and analogical reasoning enhances inferential productivity, but not representational productivity, for the output of representational redescription processes are new architectural assumptions, i.e. new representational schemes.

In a nutshell, representational redescription opens up the possibility of forging architectural constraints from the way other cognitive assets are applied, exploited or organized, rendering a *portable* representational version of the resource.

The system can them exploit the resource in new ways that were unavailable up to that point, and that makes increased portability the main goal of representational redescription.

Now that everything is sufficiently unpacked, we can see that there's no need for a ceteris paribus clause: an increasingly larger amount of representational schemes can accommodate an unbounded enhancement of the system's representational resources without losing track of what keeps each of them under control.

In other words, we don't need a ceteris paribus clause because what we want is given by the *absence* of certain possibilities in virtue of the system's architectural assumptions.

Architectural traits amounts to non-arbitrary principles that can can guide the system in its reasoning by leaving out all the permutations that are alien to that domain (things like balls falling up), and thus unless someonthing brings up very alien possibilities (relative to the domains in question), they're simply outside the system's horizon and regarded as non-ordinary.

Rather than precluding a system that relies on a single all-powerfull representational scheme from going astray, we're building up increasingly powerfull representational capacities with multiple architectural constraints.

The results allow for the right balance between the freedom to go beyond what's manifest and the world's stability.

As the system can reiteratedly redescribe its cognitive assets into representational schemes, its representational expressive power can grown indefinitely without losing track of what's ordinary within specific domains.

Getting ready for what comes next

We've handled FP at a somewhat abstract level, but is our strategy psychologically plausible?

In the next chapter, we'll see how we can go from metaphors of scale models to a plausible mechanistic implementation of structural contents.

3 THE APPLICABILITY OF REPRESENTATIONAL PLURALISM

Now that we have representational pluralism, I'll present representational cognitive pluralism (RCP): the thesis that the human cognitive machinery relies on representational pluralism by employing a plurality of structural schemes in the making up of its representational resources. In this chapter, I'll discuss the thesis' consequences and the possibilities opened up for our scientific endeavors targeting biological cognition. This requires a more fine-grained understanding of the explanatory role that structural representations can play, as well as the prospects that representational redescription gives us for dealing with higher cognitive capacities without raising the need for absolute representational contents. The result will be the outline of a set of tools placed between classic cognitivism and contemporary non-representationalism. These tools provide what's needed to approach RP without giving up on neither representations nor the methodological commitments most dear to 4E cognition.

3.1 Battle of frameworks

In the previous chapter, we've discussed representational resources from a rather abstract vantage point. As stated, one of our goals is to keep using representations even in the face of FP. But up to this point, one could still agree with pretty much all the claims regarding representational schemes made in the previous chapter and yet maintain that the discussion applies only to representations *qua* external tools, i.e. non-mental representations. There's no such thing as mental representations *qua* explanatory tools employed to understand how the mind's inner mechanisms work, but only full-blooded agents exploiting environmentally available information.[1] We use maps to drive around, music notation to learn how to play a song, graphs to express the structure of some process, and so on. This position can be found in contemporary non-representationalist approaches. They surely agree with me in claiming that there's no such thing as propositional representational contents within the mind's mechanisms, but only because they don't think there is any kind of representational content at all. Thus, if we want to make room for representations, we must show that employing multiple representational architectures is both psychologically and mechanistically plausible, in the sense that it is both compatible with current scientific practices and elucidating. That is the main target of this chapter.

The thesis that the mind employs multiple structural schemes is what I'm calling *representational cognitive pluralism* (RCP). It is basically the claim that pretty much everything previously discussed regarding structural representations can bear non-trivial explanatory roles in scientific accounts of biological cognition. Given the contemporary polemics regarding representational and non-representational approaches to cognition, RCP's position must be carefully outlined. Roughly, it is located half-way between LOT-powered classic cognitivism and frameworks fully dismissive of representations, such as ecological psychology, varieties

[1] For an example of one such account, see Carvalho; Rolla (2020b).

of enactivism and others. This is a somewhat uncomfortable position to be, for it becomes harder to make friends among both core contenders. But we can try to forge a more friendly environment for RCP by showing that its commitments (particularly those regarding representations) are compatible with the core tenets and methodological prescriptions dear to most cognitive frameworks.

Pretty much all the purchases promised by classic cognitivism relied on the hopes that a naturalistic account of representational states or processes could be provided. In its absence, the intentionality of mental states remained mysterious. To see how serious the issue is to classic approaches, one just need to remember two of its core claims. First, it conceived of cognition as essentially computing over representational states. But if one is unable to explain in virtue of what a given state can be deemed representational, the intentional character of the mind remains completely unaccounted for. It gets worst if, like most classic cognitivists and many up to this day, one adopts a semantic conception of computation. This means that even what's computational relies on what's representational. As Fodor would put it, *"no computation without representation"* (Fodor, 1980, p. 34). Second, attitude contents were a direct given of representational contents. The belief that there is beer in the fridge was explained by the existence of a representational state with that very content. Thus, if classic cognitivism can't explain how such a representational state is possible, it can't explain how a system can have beliefs, desires and so son.

There's a third core motivation for adopting representations, though it is usually underappreciated. In the 1960s the AI pioneer Herbert Simon already emphasized that complex behavior associated to rationality and intelligence did not imply proportionally complex mental mechanisms (Simon, 1996). Rather, such complex behavior could be partially explained by the complexity of the system's environment. Simon used to illustrate the idea by graphing the reasonably complex trajectory of an ant as it dealt with a bunch of obstacles. It goes around objects, it avoids steep paths, and so on. The lesson that Simon took to the heart, was that the biological machinery underlying the ant's behavior need not be that complicated, after all, thanks to environmental complexities, a relatively simple and small set of rules could result in complex trajectories.[2] For him, something along the same lines would be true of human cognition as well. Our mental mechanisms would not be simple, but they could be much less complex than we initially feared, for much of the flexibility and complexity exhibited by humans could be due to its environment. Thus, representations have a third crucial role to play: they enable us to understand how this is possible. They do mirror elements of the world, but not in any way: they do so in a proprietary format, i.e. under a particular ontology. Such a proprietary format would enable a reasonably small set of relatively simple inferential strategies to account for the flexibility and complexity we find in human behavior. The form with which we register the world (i.e. with which we encode or represent it) matter just as much as the fact that we do it.

[2] A similar remark about Simon's work can be found in Haugeland (1998d).

This is probably why classic cognitivism mitigated the importance of perception in the study of intelligence. Perceptual processes were conceived as taking place before the mind's registering of the world. Moreover, at least up to that time, they didn't seem to yield any real mystery to be solved regarding human capacities. As Smith (2019) remarked, perceptual capacities never motivated any kind of dualism throughout history. Thus, it was just plausible to assume that scientific endeavors should concentrate on how cognitive machinery registers the world. It is no coincidence that this is the space in which classic cognitivism outputs its core tenets. For instance, the idea that the mind registers the world as sets of discreet and unambiguous objects, properties and relationships, which results in a kind of absolute content and all the problems that come with it.[3] It is true that, for many at the time, the core motivation for adopting such an ontology was the naive belief that these were indeed the world's joints.[4] But that naivety is unessential and unspecific to the role that representational formats played throughout classic cognitivism's history. The fact that there were so many competing distinct formalisms and frameworks for "knowledge representation" at the time shows that the question of the proprietary form in which the mind registers the world was always regarded as a fundamental issue (Hayes, 1977; McCarthy, 1980; Minsky, 1997).

RCP is not commited to none of the two cognitivist motivations presented earlier. It does not require taking cognition to be essentially computational-representational nor posits a direct connection between attitude contents and representational contents (more on this in a bit). But it remains loyal to a portion of cognitivism's tradition inasmuch as it takes the proprietary forms in which the mind can register the world as bearing an important explanatory role to play. The way we render the world intelligible is just as important as the fact that we do it. To that extent, RCP requires a plausible theory of representational contents. This is the spot where it takes issue with non-representational accounts of cognition.

Curiously, the lesson that non-representationalists took to heart - e.g. Brooks (1991), Van Gelder (1995), Varela; Rosch; Thompson (1991) - was already present in Simon's ant. Rather than following Simon and taking the relationship between the ant and its environment as a sign that the mind works with a peculiar ontology (i.e. a certain way to register the world), they thought that we should take the ant-environment relationship as the elementary unit of analysis. The cognitive mind would no longer be conceived as a register of the world poised between perceptual inputs and behavioral outputs. It would rather be something that emerges from the continuous and direct interaction between the organism's action and perception. This is the core insight behind the idea of situated cognition.

Some cognitivists from the classic approach (or closer to it) think that the situated interpretation of Simon's ant should be fully rejected (Fodor, 2010; Sá Pereira; Souza Filho; Barcellos, 2023). Though RCP may be compatible with a classic mind-as-mirror approach, there's no reason to stick with that. Furthermore, RCP's commitment to non linguistic representational

[3] Remember the distinction between absolute and structural contents, extensively discussed throughout the previous chapter.

[4] This is the point that Hubert Dreyfus is well-known for taking issue with already in the 1960s (Dreyfus, 1992).

schemes might bother the classic cognitivist. It might see it as conceiving too much to the antirrepresentationalist. After all, the challenge of making-up attitude contents from structural representations is not really that different from the challenge of forging attitude contents from the informational structures that emerge out of the couplings and the continuous interaction between the system and its surrounding environment. In both cases, the relationship between structures and attitudes remains to be explained. This preempts claims like "to believe in p is to harbor a linguistic representation with that content". For the classic cognitivist, by giving up on linguistic representations, we could be leaving behind a crucial motivation for the adoption of representational contents. On the other extreme, a similar stance can be found: non-representationalists are wary of accepting any conception of representation, regardless of the underlying architecture and the kind of content. We can see that in their effort to argue that representations can be completely rejected in an increasingly larger set of cognitive capacities (Bruineberg; Chemero; Rietveld, 2018; Chemero, 2009; Kiverstein; Rietveld, 2018).

In the resulting picture, the first corner encompasses classic cognitivism and its refusal to accept any move involving non-representational cognition. The general worry is that it's nothing but a step back to a behaviorism-like approach (and to the kind of problems they used to face). That's why she insists in claiming that cognition is essentially related to representational contents and that this content must be, at least partially, linguistic. On the other corner, enactivists, ecological psychologists and the like are eager to fully dismiss representations in virtue of the situatedness of cognition. Their worry is that any notion of representation may be a trojan horse trying to make room for undesired methodological and theoretic commitments that brings back the problems faces by classic cognitivism.

But there are also those, like me, who think that the key is in the intersection between these two groups. This position is not really new. Its seed can already be found in connectionism (McClelland *et al.*, 1987; Rumelhart *et al.*, 1986), for this was the first major move towards rejecting linguistic representations within cognitive science. At first, connectionism was not so much about rejecting an explanatory role for representations, but rather about a promising way to register the world. This is one of the reasons why friends of connectionism usually find it easier to accept and interact with the ideas behind situated cognition. Perhaps the most well-known instance of this move is that going from Clark's *Associative engines* (1993) towards Clark's *Being there* (1997a). On the former, Clark put forward a conception of cognition that was broadly connectionist and representationalist. As for the latter, Clark has shown that his conception was compatible with some important insights coming from dynamic approaches such as that of Van Gelder (1995) and Thelen; Smith (1992). A similar - more recent - approach can be found in Piccinini (2021). The core idea is to integrate concepts from 4EA cognition without leaving aside computational and representational mechanisms. For instance, computational mechanisms are not just passive receptors of inputs. Far from it, they can be coupled with the environment and can participate in active exploration and learning processes. They can handle information from multiple sources, both from the neural apparatus (neuronal firing rates, for instance) and from the outside (visual and auditory signals).

Thus, while the antirrepresentationalism and the classic cognitivism push us towards the extremities, non-classic cognitivism push towards the center. Given that classic cognitivism is currently not taken as seriously as antirrepresentational approaches like ecological psychology or enactivism, I'll focus the discussion on the latter and try to show that structural representations can be a valuable tool even for them. Importantly, the idea is not to bring back a full-blooded conception of representationalism where every cognition is necessarily representational. Rather, what I'm commited to is the idea that representational contents can bear non-trivial explanatory roles in cognition.

In order to pay the debts I incurred in this section, I'll rely on the work of Cummins (1996). In the next section we'll discuss an important insight advanced by him: the distinction between representational contents and intentional targets.

3.2 Representations and targets

While classic cognitivism puts the whole burden of accounting for the system's intentionality in representational contents, non-representationalism puts none. In this regard, I'll side with the non-representational account. The intentionality exhibited by cognitive systems is not to be explained by the representational contents it harbors. Yet, representations are able to keep an important role in accounting for at least some of the system's overall cognitive capacities. In order to see how, we need to understand Cummins' distinction between representations and targets.

Whenever a system employs a representation as a model of something else in the world, this application can be analyzed in two distinct dimensions: a semantic and a functional one. On the one hand, the semantic dimension accounts for what a given state or process actually represents. On the other, the functional dimension is responsible for specifying the target against which the representation is applied, i.e. whatever the system is trying to exploit by means of a model. The notion of target is functional in the sense that it specifies what a given mechanism has the function of representing at any given time.

The crucial aspect of target fixation is (hopefully) non-polemical at the agent level. Say a group of friends is traveling through an unknown country at a time before the existence of Internet or smartphones. One of them is asked to provide a city-map the group needs in order to find a given touristic spot. She finds no printed map for sale, and then she starts to collect information from the locals. Based on what she learns, she draws a map herself. It includes some streets, its intersections and some presumably helpful landmarks. She meets her friends again and presents the produced map, which can be exploited to guide the group's reasoning about where to go. What's important about this example is that, whatever the structure is drawn on the map, everybody will regard it as a route leading to the desired spot. That route was her *target*. However, the produced map may not represent a proper path. It may lead to somewhere else, or maybe miss some relevant street intersection. Perhaps some local got confused or even lied to her. Thus, though it is taken to be an accurate depiction of the route towards the desired spot, the map may actually depict the route towards someplace else.

The distinction between the route actually depicted in the map and the route that the group assumed to be depicted in the map is the difference between what the map represents and the target against which the representation is being applied to. To ask for the city that a map actually depicts is different from asking for the city one is trying to think about by using that map as a model. Each of these has its own independent *explananda*. Accounting for how a system fixes on a target is distinct from explaining in virtue of what a representation has the contents it does. In other words, the capacity to have X as a target is distinct from the capacity to harbor an accurate depiction of X. That's why one can use (inadvertently or not) a map that depicts city X in order to plan a route in city Y. Moreover, while a map depicts (i.e. represents) a city at all times, targets only get fixed in particular episodes of exploitation. Thus, whenever the user of a map employs it as a tool to enable surrogative reasoning about the streets of some city, that city becomes the map's target. This means that the map's target is only determined by the system at the moment of use. The target is thus a property of representational tokens rather than types, for only tokens have their current functional roles fully determined. Consequently, theories accounting for target fixation should be distinct and independent of theories of representational contents.

The independence allows us to see that the intentionality of a given state or process is not a given of the exploited representational content. Rather, it comes from the fixed target. To fix on a target is to intend it. The capacity to intend something is remarkably distinct from the capacity to represent it. Whenever someone is asked to produce a map, she might understand what her target is, even though she doesn't have the means to hit it (perhaps she's unable to speak the local language). She's able to fix on a target that she's unable to hit. Thus, the claim that a system can target (i.e. intend) something is distinct from the claim that it can represent that target. This is an important feature of the distinction, for we don't handle all of our targets by representing them. Representational intentionality accounts for only one of the possible ways through which a system can handle targets. There can be, for instance, mechanisms specialized in detecting a state of affairs. Detectors have targets and hiccup in their presence, but they provide no information other than that, i.e. they do not rely on internal states or processes that have the function of representing those targets. We'll have more to talk about non-representational mechanisms handling targets, but for now it is important to remark that not every intentionality is representational intentionality.

From this vantage point, we can see that much of the current contemporary debate among representational and non-representational frameworks is actually about target fixation. Targets can get fixed in various ways. They may be due to the functional description of an internal mechanism, but they can also be the product of a complex process involving couplings between the system's cognitive apparatus with its body and the surrounding environment. For instance, the idea that organisms can directly exploit ecological information available in the environment is all about the targets that a system can fix on.[5] Frameworks can thus disagree about the set of targets a system can have and how exactly they get fixed. But such divergence is not about whether and how representations are exploited by the system while working with its targets. Questions about how a system fixes on a target and what cognitive resources it employs while handling that target must be kept apart. Representations are not a way to characterize that in virtue of which a system is sensitive to a range of features of the world, but rather a strategy available to the system store and exploit information about those targets.

With the distinction between target and representation in mind, we can now turn ourselves to the issues usually employed to justify pessimistic views regarding representations. There are two flanks usually exploited: the first draws on the apparent difficulty in naturalizing representational contents. The second renders representations as having no real explanatory role to play in scientific explanations, usually by claiming that semantic contents are causally inert. Both flanks lead many to reject representations and regard them as neither real nor explanatorily useful. But I think that we can overcome both difficulties. That's what we'll focus on in the next section.

3.3 What are mental representations?

Since the very beginning of cognitivism, philosophers took for themselves the task of supplying a naturalistic account of representational contents. Many non-representationalists - enactivists in particular - think that the task cannot be fulfilled, and in the last years, it became commonplace to ground this stance in the so called *hard problem of content* (HPC) as formulated by Hutto; Myin (2013). The formulation is supposed to show a fundamental and unsolvable incompatibility between naturalism and representational contents. Simply put, the core claim is that there is only one possible source of natural information, which is *covariation*, and this source won't do. Covariation is the kind of information that Dretske (1981; 1986) and others appeal to in their accounts of representational contents. Natural states or processes can covary reliably or nomically. Smoke and fire, for instance. But representationalists — the argument goes — take that to comprise a semantic relationship. Thus, whenever S covaries with F, then S is said to *mean* F. Smoke *means* fire, for they covary reliably. In the same vein, if we can find neuronal states that covary reliably with properties or states outside the neuronal system, then we can aptly call them representations.

But Hutto and Myin go on to argue that covariation cannot comprise a semantic relationship *per se*. A full-blooded agent may take smoke to mean fire because she's embedded in a social and cultural context. But no such thing can be said about sub-personal mechanisms. In other words, the semantic gloss is only in the eye of the beholder. If we take the beholder out of the picture, there's no such thing as representational relationships doing any work. There's only raw covariation. Raw covariation can participate in explanations of cognitive

[5] See, for instance, Carvalho; Rolla (2020b).

performances, but not by doing anything recognizably representational.[6] It may be relevant to characterize how the system is able to causally track something, but this non-semantic gloss is enough to exhaust covariation's explanatory role. When semantics arrive at the party, it's too late. Everything is already accounted for. That's the core issue with beholder-dependent accounts: whatever they add-up to the explanation of the performance is explanatorily irrelevant. As Dretske would summarize it: *"The fact that [representations] have a content, the fact that they have a semantic character, must be relevant to the kind of effect they produce."* (Dretske, 1991, p. 80) In this scenario, Hutto and Myin claim that friends of representations must either give up or simply hope for a rather unlikely development in physics that would - somehow - allow representations to enter the naturalistic picture.[7] If covariation cannot ground representations, and if there's no other natural source of information, representationalists should abandon any hope.

But are covariational theories really hopeless? I am very sympathetic to this line of reasoning, for I agree that one can't get symbols nor anything worth of the name "representation" out of raw covariation.[8] Of course, representationalists can always try to provide an account that grounds semantics in covariation without trivializing the semantics' explanatory role. That's pretty much what most theories of representational contents in the market set out to do. Fortunately, we can avoid that discussion. First, because non-representationalist literature is full of well-known reasons to be skeptical about this project. Even within representationalism, there are those who realize how difficult that is.[9] Second, and most important, even if somebody manages to overcome these difficulties somehow, there is an additional problem with the endeavor: theories relying on covariation entail LOT, which (as argued in chapter 2) render FP unsolvable. Let's quickly unpack this claim.

The core idea behind covariational accounts, remember, is that a certain state can be regarded as a symbol of G if, in perceptual contexts, its tokening covaries reliably with instances of G in the world. Even sophisticated accounts, such as that of Millikan (1987) and Dretske (1981; 1986) rely on this. However, in naturalistic conditions we can have only a finite amount of detectors tracking features of the world and hiccuping indicator signals whenever that feature is present. In covariational theories, such detectors are the basis on which we can ground primitive representational states (i.e. symbols). That's why we need to add composition rules in order to have representational productivity. But a system with 1) a finite set of primitives and 2) composition rules that allow it to go beyond the contents given by its primitive set is the crude definition of a LOT-powered system. As Cummins (2010c) noticed, one can surely try to dodge this simple reasoning with the following reply: there need not be a 1-1 mapping between the set of detectors available for the system and the set of primitive symbols it employs. Primitive symbols can themselves be a product of complex combinations of indicator

[6] This is what Ramsey (2007) dubbed the *job description challenge.*

[7] This claim is adopted by many proponents of enactivism, such as Rolla (2023).

[8] *Pace* Shea (2018).

[9] Extensive discussion in this regard can be found in Cummins (1996), Ramsey (2007) and Hutto; Myin (2013).

signals, which means the number of primitive symbols can be unbounded even in systems with a relatively small set of indicator signals. But what exactly would this move buy us? The result would be a rather clumsy attempt at having a symbol system that is unbounded in virtue of an open-ended set of primitives rather than its underlying combinatorics. This is empirically unsound enough, but it gets even worse when we remember that even open-ended sets of primitives have to face realistic memory limitations. In this case, such limitations will drastically mitigate the system's expressive power. In the end, this attempt to reject the claim that theories based on causal correlation imply LOT would amount to giving up on representational productivity. That's why, in a nutshell, if one is commited to a covariational theory of representation, one is commited to an LOT-like scheme. Since we already rejected the kind of content that LOT provides, HPC's success or failure against covariational theories is not really relevant for our endeavor. Even if we managed to reject HPC, we would have no interest in purchasing covariational theories.[10] To my ears the "hard problem of frame" (i.e. FP) fares much better in the market of reasons to avoid LOT.

Furthermore, whatever you may think of HPC's destructive effect on covariational theories, it cannot fulfill its wider antirrepresentationalist goals. It is simply not true that the covariation relationship is the only naturalistically plausible source of information. There's at least one additional source of natural information, and we've already met it: isomorphism. Isomorphism (or homomorphism, as "partial" isomorphisms are sometimes called) is a purely mathematical relationship between two or more structures. Therefore, even those accepting HPC's corollary can still reject the claim that HPC is fatal for those willing to make use of representational contents.

Inspired by the work of Swoyer (1991) on isomorphism, Cummins advanced the thesis that the semantic dimension of representational states or processes can be completely accounted for by this kind of relationship alone. The representation relationship is nothing but the mathematical relationship of isomorphism (Cummins, 1996). It goes for any physical state or process, from printed maps to neural activation states in the brain. A given structure A represents a given structure B inasmuch as A is isomorphic to B. Therefore, elements of A represent elements of B, relations among the elements one can find in A represent relations among the elements one finds in B, and so on. Whenever A is found, A is a potential source of information about B. We have thus a semantics for *structural representations.*

Crucially, the representation relationship does not rely on the fact that it is being currently exploited as such. We can carry with us a paper map of the street structure of Belo Horizonte (BH) all the time and never use it. The map will not become a representation of BH only when we decide to use it, for it is already isomorphic to BH even while in our pockets. Whenever we open it, it is already set up to work as a proxy, i.e. as a stand-in for the street structure of BH. Representation is, in this sense, intrinsic. There's no need for neither

[10] As we'll see in a bit, the claim is that we can't rely on any LOT-like representational theory that affords absolute contents. Covariation can keep its place in providing non-representational contents and/or working in chains of causal intermediaries within the cognitive machinery.

conventions nor covariation to play any role in determining the map's contents.

Cummins' theory is pretty straightforward, but it raises immediate doubts about how the resulting notion of content can fill the role typically attributed to mental representations. A first worry is about its seemingly excessive liberality. Isomorphic structures are too cheap and seem to be everywhere. This means that A might represent not just B, but also C, D, E and so on. Thus, a map depicting the streets of BH represents those streets, as well as the relationships among them, but it also may represent streets from other cities with similar street-structure. It may even represent any deep-ocean structure it happens to be isomorphic with by cosmic coincidence. As it stands, representational contents will always be non-unique. But this is where Cummins' distinction between target and content start to pay dividends: the pressure for content determination is actually a pressure for target determination. In other words, the non-uniqueness of representational contents is harmless. It may seem problematic, but only if we conflate two distinctive ways in which we understand utterances like "A represents B":

(1) structure A is isomorphic to structure B;
(2) structure A is exploited within a system as a stand-in or model of B.

Cummins takes the "A represents B" to mean (1). But most of the literature usually takes it to mean (2). A meaningful example of this point can be found in Hutto; Myin (2013):

> To qualify as representational, an inner state must play a special kind of role in a larger cognitive economy. Crudely, it must, so to speak, have the function of saying or indicating that things stand thus and so, and to be consumed by other systems because it says or indicates in that way. (Hutto; Myin, 2013, p. 62)

This characterization of the representational relationship is a clear instance of (2). The problem with any such instance is that it conflates the semantic and functional dimensions. What we want are the conditions under which a given state can be regarded as contentful, i.e. representational. Instead, what we get are the conditions under which we can take a given mechanism to be exploiting a state as a model or stand-in of some functionally determined target. In analogy, we ask for the city that a map represents, and we get as an answer the city that someone is trying to walk around by using the map. But the question "which city is depicted in this map?" is evidently different from the question "where are you going to use the map?". While the former is about representational contents, the latter is about target fixation. Once we have the target/representation distinction at hand, that becomes clear. But in its absence, we'll be tempted to agree with Hutto, Myin and others that there cannot be representational content unless someone's exploiting that content. This is somewhat odd, for a map of BH does not "becomes" a map of BH only when someone's exploiting it. It remains isomorphic to BH even if we lose it somewhere. The same reasoning goes for states within the cognitive machinery. Representations and targets have distinct explanatory goals. Targets are about what the sub-personal mechanism is trying to do. Whenever we say that a mechanism x has the function of representing a certain state, we're specifying its target, not its content.

That's why, unlike representational contents, targets must be unique. They'll comprise the norm against which the accuracy of the harbored representation can be measured against. That why, if some mechanism has the street-structure of BH as a target, it simply won't matter if the effectively produced representation is isomorphic to a hundred other things in the world. Mechanisms consuming the produced representation will take it to be a map of BH regardless of what the token effectively represents (i.e. of what it is isomorphic to). The representational content is not there to answer what the mechanism has the function of representing. Rather, it is there to help us understand the mechanism's performance: given that a consumer took a produced structure as a model of BH, what would be the outcome of trying to plan a route with it? Again, the fact that the produced token may be isomorphic to tons of other things unrelated to BH is irrelevant to assess its role in being exploited as a map of BH.

So far, we've seen how this approach to structural semantics avoids both HPC (for it does not rely on covariation) and the problem of non-uniqueness of representational content (it is only an issue for those who think that contents should specify targets as well). Furthermore, Cummins' structural semantics can also avoid all thee traditional issues regarding representational contents. To begin with, there's no *grounding problem*, i.e. there's no mystery about how representational contents can be anchored in the world. Grounding issues are specific to symbolic approaches. Given that the contents of primitive symbols are arbitrary references, the only way one can recognize a state as contentful (i.e. as a symbol) is by grasping that reference. Unlike isomorphism, which is grounded in the similarity of form, establishing a relationship between a symbol and its reference requires some kind of intermediary. We can easily understand how full-blooded agents can play this intermediate role by e.g. establishing conventions. But it is harder to find naturalistic and non question-begging substitutes that enables a semantic for sub-personal mechanisms. Indeed, the whole point behind HPC is to claim that there can be no such thing. This leaves us with nothing but full-blooded agents in order to establish what represents what, and that's why friends of HPC feel entitled to claim that every representational gloss will be, inevitably, in the eye of the beholder.[11] On the other hand, isomorphism need no such thing, and that's why HPC has no power over it.

In the same vein, there's no mystery regarding the causal efficacy of representational contents. Both contents and causal powers are a given of the representational token's structure. Representations are thus able to play their role in guiding cognitive processes in virtue of their contents. Furthermore, there's no need for anything like a "content interpreter". Its consumers are causally sensitive to its structure just like a lock is causally sensitive to the structure of a key. Finally, the possibility of mismatch between the structure of a target and the structure of a token enables a robust and purely semantic conception of *misrepresentation*. This last point is going to be further developed soon, for misrepresentation enables the purchase of a whole explanatory dimension that (I'll argue) cannot be purchased by other means.

Before moving on, we can quickly take stock. What do we have so far? We've seen

[11] The non-triviality of representational contents in explanations of cognitive capacities and performances will be discussed soon.

how the question of what mental representations *are* is clearly distinct from the question of how they're exploited by its consumers within the encompassing system. Representations are isomorphisms. Mental representations are isomorphisms harbored within the cognitive apparatus. Whether and how they're exploited within a system is a rather different and independent question. It involves the system's capacity to fix its targets and to employ representations of those targets whenever handling them. While theories of mental representations account for contents, theories of target fixation account for the system's targets. Indeed, as we'll see throughout the discussion, some classic theories of representational contents in the market may comprise plausible candidates for target fixation. Once they are relieved of the burden of explaining how representational error is possible, many of their best-known problems disappear. The independence between contents and targets also implies that, unlike many used to think, there's no need for a unique set of necessary and sufficient conditions to identify representational and non-representational mechanisms within cognitive systems. Rather, we must identify cases where mechanisms are exploiting structures in virtue of their form, for these are cases where a representation is being used as a model. But this is an empirical matter. One should not simply assume that, whenever there are representations, they always account for the whole performance or capacity. Their role may vary enormously, and they may be exploited alongside other cognitive resources, such as indicator signals, environmental couplings and even affective elements. Thus, in order to claim that a mental representation plays a relevant explanatory role, we must seek for cases in which they provide the best explanation of the available evidence. But how exactly should that be done? Let us now see some examples and tools.

3.3.1 Do we actually exploit structural correspondences?

The human cognitive machinery is clearly capable of exploiting causal correlations through indicator signals coming from detectors. A detector is basically a mechanism that hiccups whenever some target state of affairs obtains. There's plenty of well-known examples of the crucial idea out there: thermostats indicate the ambient temperature; idiot lights in cars indicate whether it's time to get more gas; predator detectors enable non-human animals to run away, and so on. As the work of Hubel; Wiesel (1962) established long ago, many cells in primary visual cortex (V1) can be regarded as edge detectors, and the electrical spikes emitted in the presence of such edges is their signal. Importantly, indicator signals are not representations. Their content comes from the functional specification of the detection machinery. The yellow light of a car indicates low gas because of its place and function in the car system, not because the yellow light is a symbol. The information conveyed by the yellow light depends on its source. That's why you can't tell the meaning of two yellow lights in two distinct systems without looking at their functional specification. The same goes for the signals coming from V1's edge detectors. In order to behave like symbols and constitute something akin to LOT, indicator signals must be made portable, i.e. there has to be something in the signal itself that mitigates its source-dependence and allows one to focus solely on the signal's physical traits. Grounding symbolic contents in indicators is the idea that HPC tries to gloss as impossible. But though we should be wary of symbolic machinery, it seems hard to deny that human cognitive mechanisms exploit indicator signals.

In contrast, the idea that we exploit structural correspondences may not be so readily acceptable. The mere existence of structural similarities in the brain's activation patterns and wordly structures is not enough to show that it's the similarity that's being exploited. The processing cannot be neither assimilated by nor reducible to the exploitation of causal correlations. That's easier said than done, for the relationship between signals and structures can be quite complex. In particular, structural representations might be constructed from indicators signals and yet their contents might be disjoint. Let us start with an example of how that complexity can come about in image processing.

3.3.1.1 First example: natural images

The research of Olshausen; Field (1996) targeted the brain's processing of natural images and established what came to be known as the *sparse coding* generative model.[12] The targeted images involve natural scenarios (mountains, trees, lakes, etc.) with no human-made artifacts, such as cups or buildings. A remarkable feature of natural images is in their statistical structure. As Olshausen and Field point out:

> The activities of the photoreceptors themselves do not form a particularly useful signal to the organism because the structure present in the world is not made explicit, but rather is embedded in the form of complex statistical dependencies or redundancies, among photoreceptor activities. (Olshausen; Field, 1997, p. 3311)

This lead many to speculate about what exactly would be the task of the visual system in handling the statistical complexity found in photoreceptors. What is it doing with them? Barlow's early suggestion was that visual systems have the goal of extracting these statistical dependencies and decorrelating them (Barlow, 1961, 1989), generating a collection of maximally statistically-independent descriptive functions. Natural images would thus amount to collections of maximally independent indicator signals from such basis functions. The crucial idea behind the pursuit of such independent indicators is that of reducing the redundancy found in photoreceptors. For instance, since there's a lot of spatially continuous stuff in natural images, a certain light incidence over a single photoreceptor is usually a sign that its adjacent region have similar light intensity and color. This is a way to capture our intuition that images are comprised of structural primitives, such as edges, straight lines or curves. The point brought forth by Olshausen and Field is that the set of maximally independent basis functions would constitute a kind of sparse code in which the core features comprising the picture can be captured.

[12] My attention to this work was drawn by Cummins; Poirier (2010) and the reasoning I present in this section owes much to the one they've developed there. See also Olshausen; Field (1997) and Olshausen; Field (2000).

Importantly though, this is not a general principle for image processing (Olshausen; Field, 1997), but a kind of empirical thesis. One can only decorrelate the statistical structure and find its maximally independent descriptive primitives if they're actually there. With this in mind, they designed an algorithm to figure out the likely set of maximally independent basis functions, that is, they managed to reach a candidate to the structural scheme applied in constructing natural images (Olshausen; Field, 2000). The number of basis functions they've found is remarkably small, and the set is also notably similar to the edge detectors mapped in V1 by Hubel; Wiesel (1962) a long time ago.

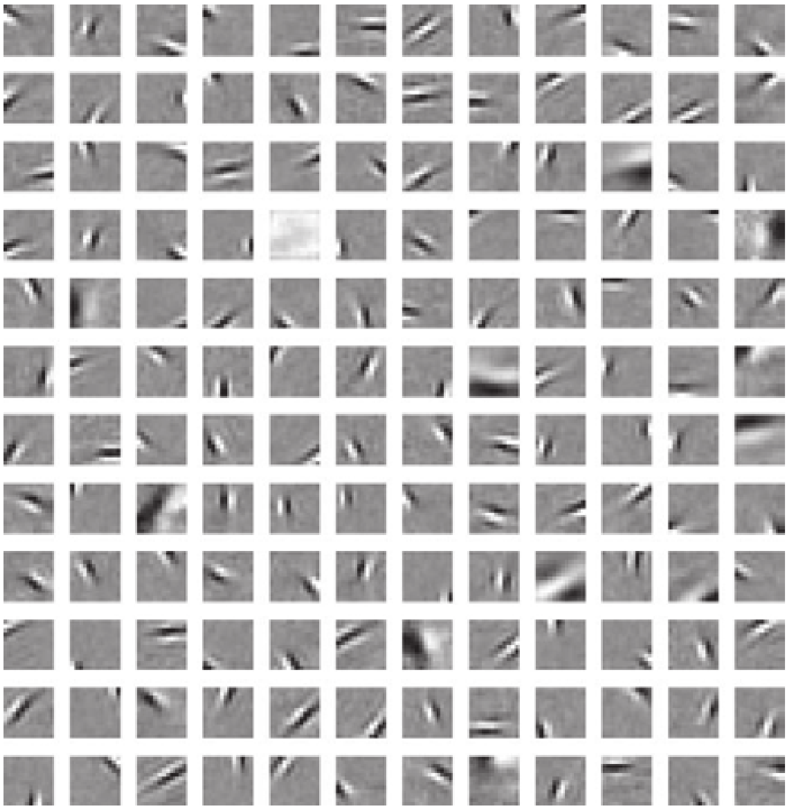

Figure 8 – The smallest set of maximally independent basis functions for constructing natural images. [13]

The basis functions express the brain's sensitivity to certain visual features of the environment, that is, they are a kind of indicator signal. Thus, it is tempting to ask whether a visual input can be a composition built directly from the information they carry. That would amount to an explanation of natural images in terms of indicator signals pointing to environmental properties: the image's content would be a product of indicator's content. However, this works with the assumption that each indicator is completely independent of each other rather than just maximally independent, and this assumption is false. Here's the author's warn against this temptation:

> Typical images are not simply the sum of light rays coming from different objects. Rather, images are complicated by the effects of occlusion and by vari-

[13] Picture extracted from Olshausen; Field (2000).

> ations in appearance that arise from changes in illumination and viewpoint. What is more, there are often loose correlations between features within a single object (say, the parts of a face) and between separate objects (chairs, for example, often appear near tables), and independent-components analysis would erroneously consider such objects to be independent entities. So the most one can hope to achieve with this strategy is to find descriptive functions that are as statistically independent as possible. But it is quite unlikely that such functions will be truly independent. (Olshausen; Field, 2000, p. 244)

The basis functions express the brain's sensitivity to certain visual features of the environment, that is, they are a kind of indicator signal. By pointing their status as not completely independent of one another, what the authors are saying is that indicators need help in detecting their targets. This is no threat for the sparse code approach in general, for one of the motivations behind the pursuit of maximally independent descriptive functions is that of explaining the brain's efficiency in processing natural images. Thus, maximizing redundancy avoidance is a possible way to achieve increased efficiency. But images cannot be simple collections of independent detector signals, for the same set of descriptive primitives is able to account for distinct images, and we can't tell them apart just by looking at the collection involved in its construction. The geometrical structure in which those signals are arranged is also necessary. As an illustration of this point, consider again the scheme Palette-C, discussed in the previous chapter.

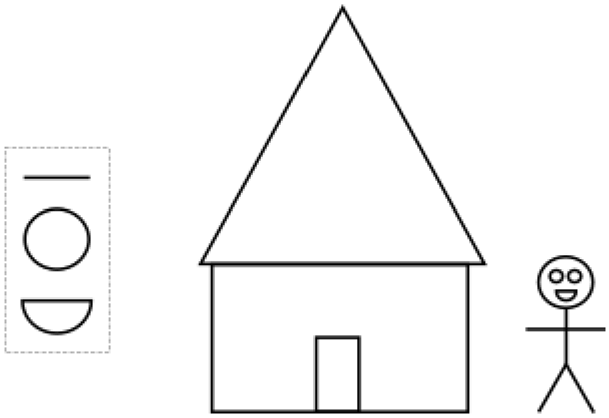

Figure 9 – The Palette-C scheme, again.

The thing about Palette-C is that its primitive constituent's meaning is completely disjoint from the meaning of its tokens. The scheme's minimal semantic element is not identical to the set of its basic semantic elements. We can surely claim that the meaning of the first primitive is a straight line, but it'd be pointless to ask whether the meaning of that straight line is being a human arm or the roof of a house. We don't know that except within an already worked out representation. This is why, if we try to account for the contents of figure 9 in terms of information yielded by a constituting collection of indicator signals, we would need something like a house-signal and a person-signal. But that would fall short of what we need, for the house includes a door, and the person has a face (she's smiling). These contents are there in the token for us to exploit, but if we adopt a single "there's a house here" signal, that information will be lost. This is the price for making the basic elements completely independent of one another: the resulting scheme becomes language-like and, as we have seen when discussing the incomensurability of distinct schemes, its contents would be absolute instead of relative. Olshausen's point about true independence as not achievable entails a scheme that is not language-like.

If the basis functions, understood as indicator signals, are not enough to explain the contents of natural images, then what can account for the scheme's geometrical properties? There has to be something else involved. A suggestion of what this can be is found in Lewicki; Olshausen (1999). They add a Bayesian gloss to the image processing mechanisms by claiming that indicator signals rely on probabilistic information about the visual surroundings. As a consequence, indicator signals must rely on a lot of background cognitive knowledge about the visual environment. In the resulting picture, indicators do have a clear role to play in the construction of a natural image, but this role is not that of providing the basic semantic elements out of which an image can be constructed. Rather, they allow for the efficient processing of statistic structure.

That's just the tip of the iceberg, though. The sparse coding model prompted a whole field of study. An increasing number of researchers started to propose refinements and expansions, both in how to better capture the statistical structure of natural images and in finding out how could a neural substrate implement it. As examples of the former, there's the approach of Karklin; Lewicki (2008) to reach higher-order statistical structure. As for the latter, there's the work of Boerlin; Denève (2011) suggesting that recurrent neural-networks provide an adequate architecture for implementing sparse coding strategies, and there's the concurrent approach of Gregor; LeCun (2010), for whom feed-forward models are better suited.

Furthermore, the approach was applied to other modalities. Though the early work focused on the spatial dimension of static images, the research of Hateren; Ruderman (1998) made similar findings employing the same kind of analysis on video sequences of natural scenes. Just like the basis functions found by Olshausen resemble edges, van Hateren's functions resemble shifting edges. They managed to obtain a similar set of maximally independent functions that shift with time. Later on, Caldieu and Olshausen would also apply a similar approach in order to capture motion (Caldieu; Olshausen, 2008). The work on motion is specially remarkable for us because it's an example of how we can represent the dynamics of a domain, a possibility that, as we have seen in the previous chapter, is unavailable for language-like schemes.

Needless to say, these examples are not supposed to express the bleeding edge research nor the current state-of-the-art in image processing. The point is just to provide a small glimpse of how fruitful and robust the sparse coding approach is and thus justify its use in our discussion about structural schemes in the human mind. In the end, we have structural representations of natural images whose content is distinct from the information brought about by sets of indicators. Just like palette-C has no indicator for arms, legs or house-parts, there's no such thing as indicators for the visual properties specific to rocks, trees or mountains, i.e. the contents of natural images are not due to the existence of indicators for rock-edges, tree-textures or mountain-shapes. The image's content is a product of the token's geometry rather than the source of the involved signals. Once a structural scheme is stabilized, the system is free to exploit its systematic permutations. One can, for instance, render the human in the picture as being sad by rotating the semicircle representing its mouth 180 degrees. Or one can also try new compositions of the primitive elements, such as using a semi-circle and a straight line to draw a hat over the little person's head. As discussed in chapter 2, these are the kind of systematic permutations afforded by structural schemes.

3.3.1.2 Second example: rats, mazes and food

The motivation for providing additional examples is that the existence of structural similarities between neural states and things out there does not imply that those similarities are being exploited in virtue of the structure shared with the environment. Indeed, the study of natural image processing has shown that we might need a considerable amount of work before claiming that a given capacity is better explained by exploiting structures rather than, say, by relying on sets of indicator signals. Importantly, this is not just about contrasting cases of representation exploitation and indicator exploitation. We'll see how the same idea can be applied to situations in which we must determine whether some cognitive capacity or performance involves representations, direct sensitivity to affordances (as emphasized by ecological psychologists) and even the active exploration of the environment by the system (as emphasized by enactivists). In order to reinforce the idea that structure exploitation is a feature of biological cognition in at least some cases, we need to show that and how structural representations can also participate in non-visual domains.

Cognitive maps are perhaps the most used example of structural representations in biological systems. That doesn't mean they're not polemical, though. Such maps are supposed to account for our capacity to go from a point to another even without any external signals pointing towards the right direction. The firing of so called "place cells" in the rats hippocampus, for instance, is know to be sensitive to the rat's location in space. That's enough to enable simple associations such as "doing action X here will be rewarded with Y". Associations like that can be explained using indicator signals. However, it has also been noticed that the relationship between co-activation patterns in place cells structurally resembles spatial locations (Burgess; O'Keefe; Recce, 1993; OKeefe; Burgess, 2005). The firing of a place cell increases the likelihood of firing other place cells that are sensitive to the animal's spatial surroundings. This is built up from both exploration and previous cognitive knowledge. The resulting structure of co-activation can be used as a proxy for the spatial structure in the rat's environment. It enables the rat to exploit its structure in order to plan ahead routes towards pre-established locations, and there is evidence that's exactly what they do (Pfeiffer; Foster, 2013). Therefore, the evidence suggests that the rat's capacity is not explained solely by indicator signals. Rather, they manage to go from a know place to another through the exploitation of geometrical structure, just like we do when using a printed map.

All the above would be enough, provided that representations were the only possibility. However, enactivists can insist that the rat is just actively exploring its environment and relying on non-representational affordances. Thus, it's not enough to show that some explanation involving representations is available. We have to show that the representational approach is the best available explanation, considering the evidence. An interesting example can be found in Newen; Vosgerau (2020). The authors emphasize the rat's capacity to articulate information regarding not only the spatial features of the maze, but food and time periods as well. They can understand that, if a certain kind of tasty food was left in a certain position of the maze in the morning, it means that there will be another kind of tasty food (say sweet) in another position of the maze in the afternoon (Crystal, 2013; Panoz-Brown *et al.*, 2016). For Newen and Vosgerau, the best available explanation to this performance involves articulating representations of food type, day periods and spatial location. In particular, they think that this example is specially useful to show how a non-representational approach amount to a poor fit of the data:

> The informational state of rats which have learned to behave according to a conditional in the maze is best characterized as structured into components of <object-type; location; time>. The alternative would be to presuppose a high number of independent, non-structured dispositions which need to include all the possible permutations of associations between a starting state of affairs and a type of behavior. And these dispositions would need to be learned independently of each other, since there would be no common component to be taken over. (Newen; Vosgerau, 2020, p. 184)

Though it is conceivable that the rat's behavior can be explained by a reasonably big set of distinct affordances, this alternative would require at least one independent affordance for each exploitable permutation of the represented structure. Otherwise, the rat's performance would remain partially unexplained. This is obviously no problem for the thesis that there's ecological information available, nor for the claim that they can play important roles in the explanation of some cognitive capacities. The issue is that, in this case, in contrast with the representational approach, appeal to affordances and active exploration don't fit the available evidence so well. The authors point out that the rats present a quick and flexible learning rate. This is specially true in cases of conservative permutations of situations the rat is already familiar with. For instance, they can understand that the maze was partially rearticulated, or that, in some conditions, a whole maze sector can be considered irrelevant to determine how likely it is to find food at night. Accounting for this kind of flexibility with a non-structured set of affordances would imply a rather distinct learning trajectory, possibly slower and more error-prone. But the point is not that representations allow the rats to make fewer mistakes. Rather, the point is that the kind of behavioral output they exhibit throughout the learning process would be different, specially in cases requiring complex permutations involving the elements of the triad object-place-time. The upshot is that, in at least some cases, representational contents may provide the best explanation of the available data.

3.3.1.3 Third example: causal structures

The examples previously exploited comes from distinct domains. Let us now consider the case of causal structures. Evidence shows that 3-5 years old children are already sensitive to the underlying causal structure of fictional stories (Walker; Gopnik; Ganea, 2014). This allows them to generalize some of the story's contents to the real world, depending on how similar they are (fantastic worlds decrease the likelihood of doing so). Furthermore, there is evidence that 7 years old children are already fully able to recognize the same underlying causal structure underpinning distinct events in distinct stories (Rett *et al.*, 2021). The capacity seems to be already present (though in a less accurate version) in 5-6 years old children, where the researchers found mixed results. Structural representations fit the evidence nicely, for they allow us to understand those capacities as an articulation of how accurately a given structure is captured and how skillfully is it exploited. Just like the previous examples, that's an empirical matter, so if we can find evidence of at least one case in which causal reasoning must rely on the exploitation of structural correspondence, the suggestion becomes even more appealing.

A nice example can be found in Huys *et al.* (2012).[14] The researchers formulated an experiment in which the subject's behavior relies on the exploitation of causal structures. That was not the main goal of their research, though. Rather, their target was determining the planning strategies that subjects employed while assessing their choices (i.e. the heuristics or hints they were using in trying to maximize their reward). In doing so, the researchers have shown that the subjects were evaluating sets of possible paths afforded by a causal structure. The strategy employed was akin to the ones we would use in assessing the shortest path out of a maze. This is how the researchers depicted the causal structure employed in the experiment:

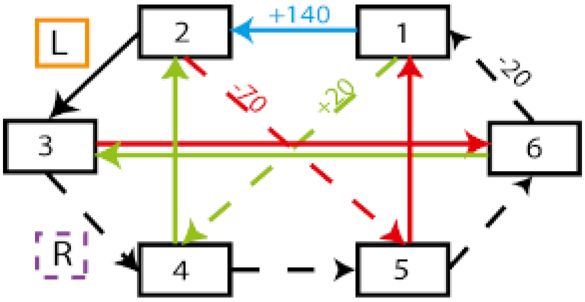

Figure 10 – The causal structure employed in the experiment.[15]

Here's a glimpse of how the experiment was performed. Subjects were presented with a screen depicting the six available locations, and one of them was marked as their initial state. All they could see was a set of rectangles, that is, they had no access to the connections depicted in figure 10. Every location affords two distinct moves, and each move can provide either a reward or a cost in pences. By convention, the possible moves were distinguished as left ones (L) and right ones (R). Thus, moving through the structure could be described as

[14] The work of Huys et al. was brought to my attention by Shea (2018). See also Huys *et al.* (2015).

[15] The picture was extracted from Huys *et al.* (2015).

sets of left or right moves such as "left, then right, then left again". But those left/right moves are not to be taken literally. They're not a function of the visual depiction of the state. The resultant state for left moves are depicted in dashed arrows, while the consequences of moving to the right are expressed with solid arrows. For instance, if the subject is in state 4 and moves to the left, she'll be taken to state 5 rather than 3. Furthermore, each transition could result in a loss or a reward. The blue arrow brings with it the largest gain (+140), black arrows result in a small loss (-20), while red arrows result in larger losses (-70). Finally, green arrows result in a modest gain of +20.[16] But not even this set of single-move consequences was available for subjects. They were asked to provide sequences of three to five left/right moves and the only feedback they would get was their final state and the final sum of their reward or loss. For instance, if you happen to be at location 3 and tries a left-right-left sequence, the feedback would be a reward of +50 and the final state would be 2. Like rats in a maze, the subject must manage to learn the underlying structure by trying distinct combinations of moves and assessing the aggregated results. In doing so, they could learn the maze-like causal structure of the task, and exploit it to plan ahead and try to get the most rewarding final consequence.

Depending on the number of guess moves, the set of possible sequences increases quickly. Since each state denotes a tuple of two distinct consequences (left, right), there are 8 possible 3-move sequences, 16 possible 4-move sequences, 32 possible 5-move sequences, and so on. The researchers have shown that some subjects would use heuristic strategies that quickly dismiss paths involving a major loss (-70) even if that path would eventually lead to a positive outcome. In virtue of such strategies, those paths were not assessed in their entirety. The relevant point for our current interest is that this kind of reasoning relies on some kind of representation of the relationship between the six possible states. The fact that we can employ distinct strategies (whatever they are) to think ahead while handling the same structure strongly suggests that we are exploiting an internal correspondent structure as a proxy. This is an example of surrogative reasoning about a causal domain. The same strategy employed by the rat's hippocampal place cells can be applied here: we can learn to navigate throughout the causal structure of the task as we learn to navigate a maze by trying out distinct paths and assessing their outcomes.

3.3.2 How to tell whether one is exploiting representations?

This small set of examples is not supposed to provide evidence that every cognition involves the exploitation of structural representations. This is not a return to representationalism as the idea that every instance of cognitive processing is necessarily an instance of representation-guided processing. Rather, the claim is that sometimes we can (and do) ex-

[16] If the reader is relying on a scheme that can't represent color (such as those used in black-and-white printings), that will result in a forced representational error, for the blue, red and green arrows will be necessarily depicted as distinct shades of gray. In this case, the following language-like encoding might be helpful: the blue arrow (+140) is the one depicting the transition 1-to-2; red arrows (-70) are depicted in the transitions 3-to-6, 2-to-5 and 5-to-1; finally, green arrows (+20) are depicted in the transitions 1-to-4, 4-to-2 and 6-to-3.

ploit structural correspondence for distinct purposes in distinct domains. Finding out when we actually do that is a purely empirical matter, and the capacity in question or performance must be assessed in all its specificity. Thus, it's useful to briefly mention the kind of strategy that allows us to confirm whether a given capacity comes from the exploitation of structural correspondence or something else. How can we distinguish cases where the system effectively exploits representations from those where it is exploiting other kinds of resources, such as indicator signals' environmental couplings and the like?

Merely detecting instances of structures within the cognitive machinery won't do. Even if the representational structure is there, this tells us nothing about the role it is playing in the cognitive economy. It may be exploited only as an indicador signal, i.e. the system may ignore its form and simply take it to be a sign of something else. Indicator signals may be complex in form, but their content is not related to their formal properties. One could easily put a complex light pattern in a vehicle and claim that, whenever that pattern obtains, it means the vehicle is running out of gas. The light pattern is there and, being a structure, it can definitely represent something with that form (maybe the lights have the shape of a dog), but this is completely ignored by its consumer, who takes the token to mean low gas, and nothing more. In a similar vein, Facchin (2021) noticed that detector mechanisms can render states isomorphic to the kind of feature they target in the world. It is worth to see his reasoning in full:

> Consider (...) the bi-metallic strip of the thermostat. Let it be sensitive to three environmental temperatures, ordered by hotter than in the triplet (t$_a$,t$_b$,t$_c$). Let v$_a$, v$_b$ and v$_c$ be the corresponding states of the bi-metallic strip. Suppose now that longer than orders these states in the triplet (v$_b$,v$_c$,v$_a$), preventing the relevant strip-temperature structural similarity from obtaining. Yet the strip can still successfully orchestrate the behavior of the thermostat, at least as long as it enters in each state when the environment is in the corresponding temperature (i.e. as long it correctly indicates) and each state leads the system to behave as it has been designed to behave. So the relations among the features of the vehicle are irrelevant to the functioning of the system. As a consequence, the structural similarity is not exploited, as a structural similarity is exploited only if a system is sensitive to the relations among the features of the vehicle (...). Receptors might be structurally similar to their targets (and as a matter of fact they are). Yet, this similarity does nothing for the system and deserves to be called an epiphenomenon (...).

Facchin's reasoning generalizes, meaning that every indicator mechanism is isomorphic to some structure of the target feature. He's obviously right in claiming that the system is not causally sensitive to that structure, i.e. the structure is not playing any cognitive role in virtue of its form. Facchin thinks that the structural similarity is thus epiphenomenical. That's not quite right, though. An epiphenomenon would never be able to play any cognitive role, but that's not true of those structures. First, like any other structure, they represent what they're isomorphic with. This means that, despite being non-exploited, they're not devoid of their representational status. They're instances of what Cummins dubbed non-exploited representational content (2010). Non-exploited contents are simply cases in which representation-producing mechanisms token structures that are not exploited or that are only partially exploited by consumer mechanisms. The structure of indicator mechanisms is just one such case. Therefore, it is possible, at least in principle, that the system can learn to exploit those structures in virtue of their form. This is enough to reject the claim that those structures are epiphenomenical. They do represent, and the issue whether a particular cognitive system can learn to exploit such structures is empirical.

In standard representational stories, this would seem to be, at best, a quite peculiar situation, but that's actually the rule rather than the exception. One cannot exploit representations that are not there to be exploited in the first place. As Cummins would say: *"You cannot have a representation drive story about learning if your story about representational content implies that the content that is supposed to do the driving is only there when the learning is complete."* (Cummins, 2000, p. 117). We can see this point at work in contemporary machine learning techniques. Large language models such as those of the GPT family take their training *corpus* to represent a structure of probabilistic distribution that describes where and when a given syntactic element is expected. As the training goes by, it learns how to exploit that structure with increased accuracy. If we pause the learning process at any point, the result will be a consumer that is only partially sensitive to the structural information available in the representations. Something along the same lines can happen in the brain. Visual subsystems, for instance, may represent (i.e. produce isomorphic tokens) of the proximal *stimuli.* Such representation may already encompass structural information about deepness, thus allowing consumer mechanisms to infer and exploit it. But there may be no consumer mechanism able to do that. Whether the system can eventually become sensitive to it depends on its set of available learning strategies or on its overall developmental trajectory.

A similar reasoning goes for the structures involved in receptor mechanisms. Rather than just providing information about the presence of some feature ("feature x here now"), detector mechanisms are a potential source of structural information about their target features. If something like this actually happens, it amounts to a substantial change in the functional profile of the detector in question. While detection information is source-dependent, the structural information it provides about its target feature is not. Even if structural information is "copied away" from the detector mechanism and employed someplace else, its representational content remains the same. Source-independency is a kind of decoupling that's unavailable to non-representational assets.

The upshot is that the mere presence of a structure is enough for it to be considered representational. It is not enough, though, to tell us that the structure is being exploited in virtue of its form. How are we suppose to distinguish such cases? Shea (2018) suggests that we should investigate the correlation between degrees of accuracy and behavioral outcomes: how would the system perform in the presence of a less (or more) accurate model its target? The underlying assumption is that whenever improving the accuracy of a given structure increases the likelihood of good performance, then it is likely that the performance is achieved by the exploitation of that structure. But without further development, the suggestion leaves an important feature either mitigated or completely out of the picture: semantic success does not imply functional nor behavioral success. Good performance may come with bad accuracy, and bad performance can occur even in the presence of accurate structural resemblance (we can make the most out of bad maps, and we can fail to correctly exploit what's given in very accurate ones). Thus, the pursuit of evidence for the exploitation of structures should not work under the assumption that good accuracy implies good performance. Predator detection, for instance, is probably better accounted in terms of indicator signals in most creatures. But just for the sake of the example, suppose there's a creature that relies on the structural resemblance of its know predators to predict their moves and decide whether they stay or flee. Mechanisms with lots of false-positives could be adaptive even if the resemblance was loose, or even if the exploitation was error-prone. Greater accuracy usually imply more cognitive resources than lower accuracy, and predator-detection systems are specially sensitive to this: the quicker and cheaper, the better. Importantly, the answer to the question "what's the proper level of accuracy?" is clearly dependent on the present circumstances. In some situations, there's room for extensive use of time and memory, and more accurate representations would provide benefits. In others, there can be serious time and memory constraints (the organism is trying to outrun a predator), and these might affect the proper level of accuracy, i.e. the level providing the best trade-off between precision and speed.

The distinction between good representational accuracy and good performance is one of the main purchases of introducing representations as an explanatory tool. Though it might make things a bit more complicated, we get in return an important degree of articulation. One can articulate distinct empirical hypotheses regarding the mechanism's target, the produced representational content (if any), how the content is exploited (if it is) and what's the resulting behavior in distinct circumstances. What we need is a way to empirically determine the overall strategy that's being employed by the system in exercising a given capacity. Assuming that we can see or measure what's going on in a given mechanism, how could we draw any conclusion about the form of the mental representations being exploited there? A nice suggestion can be found in Blackmon *et al.* (2001): we should look for *incidental effects.*[17]

Say we're trying to figure out how a given calculator can compute the product of 25 times 5. We have two distinct candidate models: the first claims that the inner processing occurs by successive sums. The calculator goes on to accumulate the partial addition of 25+25+25+25+25 and then present the final result. The second model suggests that the inner processing relies on the partial products algorithm. Thus, the system would break the multiplication into smaller steps (5.5, 5.2...), and then proceed to aggregate the partial results. Both models account nicely for the calculator's behavioral output (printing 125 on screen). They are, to use Pylyshyn's term, weakly equivalent (Pylyshyn, 1984). But these models differ in other aspects. Successive sums render the system sensitive to how big the multiplier is: it will take longer to multiply 25.9, for 9 is larger than 5. In its turn, the partial products algorithm

[17] See also Cummins (1983) and Cummins (2010a).

makes the system sensitive to the number of digits of the multiplier: 25.5 and 25.9 will take the same amount of time. Moreover, though increasing the number of digits in the multiplier (say 25.12) will increase the amount of time in both models, it happens at distinct rates. This temporal effect is an example of incidental effect. It is incidental in the sense that it is secondary: it won't change the input-output relationship. Lots of things can be a source of incidental effects: environmental interference, unusual working circumstances (e.g. high temperature), the processing speed, and so on. In this case, the effect is a function of the algorithm employed by the calculator. If we can insulate them and measure the algorithmic effects, then we can use the data to confirm which model better accounts for the calculator's capacity.

A real world application of this strategy can be found in the example of how rats exploit structural representations. The argument from Newen and Vosgerau (2020) appeals crucially to incidental effects. More precisely, they claim that, even though the kind of behavior can be explained without representations, this alternative would imply incidental effects (e.g. a distinct learning rate) that are incompatible with the current set of available evidence.

A similar reasoning can be used to distinguish cases in which the cognitive apparatus is relying on representations from those in which it relies on encodings. Different schemes provide distinct *systematicity effects*. A systematicity effect is one that implies sensitivity to the structure of some domain. Natural images and causal reasoning were all examples of measurable sensitivity to the systematicity of their respective domains. In chapter 2 we have seen that representational schemes can exhibit systematicity effects in two ways: by representing the domain's structure or by encoding it. For instance, language-like and structural schemes can exhibit distinct temporal profiles when parsing linguistic inputs. Long ago, Smolensky and others have shown that neural models relying on non-linguistic schemes can be weakly equivalent to language-like models in their capacity to parse the structure of some linguistic input (Smolensky, 1990; Smolensky; Legendre; Miyata, 1992). In order to establish that, they used tensor products to encode linguistic domains. Unlike classical language-like schemes, tensor products do not resemble the systematicity of linguistic domains in any way. Such models can handle an entire sentence in a single step, for the many nodes comprising a hidden layer can work in parallel.[18] On the other hand, linguistic-scheme-powered models might need to analise distinct parts of each sentence, one by one. The temporal signature of these models will be distinct, and this kind of distinction allows us to confirm which one better fits its target. Thus, systematicity effects can be used as incidental effects to decide for the best available model.

Going further, as Blackmon *et al.* (2001) notices, distinct models describing the way human cognitive machinery parse linguistic inputs may also have distinct *priming effects*. Language-like schemes, for instance, might predict that the closest systematic variants should be more easily accessible. Thus, we could investigate whether an input like "John loves Mary"

[18] This is not exactly true about artificial neural networks, for they usually run in serial processors. But that's not an issue, for the temporal signature will still be distinct. The time it takes to process a hidden layer won't change as a function of the input size, for instance.

primes the mind in a way that allows it to process anything involving those elements in a faster way. If the systematic linguistic variants of a given input are more easily accessed after its processing, that can be regarded as a sign that there's a language-like scheme underneath. On the other hand, priming effects that are inconsistent with the systematicity of language-like schemes can point us towards a non-linguistic scheme.

As a final quick example of how incidental effects can be informative, we can also consider what happens when they're lacking. The work of Cao (2020) regarding predictive processing and the interpretation of neural signals is a nice example. She makes a point about how generative models (the kind typically associated with the predictive processing framework) and non-generative ones can be regarded as mathematically equivalent. What this means is that we're not aware of any incidental effect we can use to decide between those (at least so far). If the absence of incidental effects persists through time, it might be suggestive that, on the one hand, we're on the right path towards accounting for cognition as it really is, for both are capturing something true of cognition. But on the other hand, it also means that there's no new game in town, i.e. predictive processing could be, as Cao suggests, a new gloss for old ideas. What matters for us is that the confirmation of whatever turns out to be true relies on incidental effects. In order to prove that Cao is in error, one would have to come up with a plausible candidate for incidental effect, and maybe an experiment that confirms its presence.

3.3.3 The role of representational contents in the assembly of attitude contents

We've just seen some evidence that humans exploit structural resemblance in at least some occasions. But how much of cognition can be accounted for in this way? And how can those structures give rise to attitude contents such as beliefs and desires? Both questions are empirical and require extensive scientific research. Whether most or few of our cognitive processing relies on structural representations will come up from the accumulated body of scientific knowledge. The same goes for how structures engage or participate in attitude contents. But the latter claim is commited to a somewhat radical (and not always appreciated) departure from the way classic cognitivism thinks about representations.

When Fodor's LOT reigned, much of the literature used to think about the relationship between representational contents and attitude contents through the lens of *Fodor's dictum*: to believe in p is to harbor a representation with the content p in the "belief box". A belief box is just a characterization of the role played by a token at a given task. In this case, the role is akin to that of taking something as true. Consequently, the dictum puts all the explanatory burden on representational contents. Attitude content is nothing but representational content. Therefore, to put forth a theory of attitude contents comes down to providing a theory of representational contents. Given the dictum, the resilient preference for language-like schemes is no surprise. How could we account for propositional attitudes without language-like representational contents? The impression that there's no plausible answer to that question is behind much of classical criticism against structural representations. How should we "picture" logical properties such as negation or disjunction?[19] What could be the structure of normative properties such as justice or good? Fodor's dictum lies behind such questions, for they work with the assumption that attitude contents must be somehow reducible to (or completely accounted for in terms of) representational contents.

In order to grasp the explanatory role of structural contents, we need to introduce the notion of *application*. Classic work in mental representation conventionally expresses a token whose content is x as "|x|". However, the notation is ambiguous with respect to the target and content dimensions. Assertions such as "the system tokens an |x|" are used in order to express both that the system harbors a representational state with the content x (content dimension) and that the system intends x, i.e. that the representation is supposed to be about x (target dimension). In order to disambiguate it, we can follow Cummins' jargon: whenever a representational mechanism exercises the function of harboring a token of x, we can say that the mechanism is *applying* a representational token to its target: $|\mathrm{x}| \rightarrow x$. Applications are distinctive in the sense that they have both dimensions (target and content) fully established.

Another difference is that, in classic approaches, when a system tokens an |x|, we're entitled to infer its contents: x. However, when it comes to applications, this apparently straightforward inference is not available. Application contents are distinct from representational contents. This provides an extra degree of freedom for scientific theories about cognitive mechanisms, but unpacking applications gets trickier, for we need to know both content and target, and this is not always obvious. To see why, let us consider first an example were both target and content are know in advance. Say a system harbors an application like |board| → *current_board* in order to keep track of a chess game. The fact that the pictured board is the one from the current game is nowhere represented on the system, i.e. it cannot be found in |board|. The information is available, but only because it can be inferred from the functional specification of the mechanism doing the targeting. As far as the contents of |board| go, it makes no difference whether the pictured structure is that of the current game being played by the system, a hypothetical position the game would be in after a move being considered, or even the status after the 27h move of last year's championship final. Therefore, the content of an application is not identical to the representational content being applied. The content of |board| is simply a board structure, but the content of its application by a mechanism with the function of representing the actual board of the game currently being played is something like *"the current board of the current game is |board|"*.

On this account, the explanatory role played by representations in cognitive systems can only be fulfilled by *applications*. Targets to which no representational content is being applied are not representational assets. Representational contents that are not being applied to a target cannot account for representational error. We have thus a notion of misrepresentation that only requires applications. But what is the relationship between applications and attitude contents? We could advance a theory of attitude contents that's as simple as it gets: an attitude

[19] It is not new nor surprising that perhaps some logical properties can indeed be pictured. Shin (1994) set himself the task of showing how this could be done with Venn diagrams.

is simply an application with a cognitive role. All we'd need to do is give the application a cognitive role, such as that of being a belief, a desire or a goal. Couldn't "the current board of the current game is |board|" work as a belief for the system? Though that's possible, it would amount to Fodor's dictum 2.0. Once we realize that representations need not account for everything there is to handle regarding attitudes (as we just did by seeing that the content of an application is also given by its target), why should we restrain ourselves like that by assuming that attitudes are the sole product of applications?

Attitudes can be much more complex and comprise articulations of multiple applications, attention mechanisms, affections, indicator signals and other kinds of cognitive resources. Some of these resources may bear closer relations to attitude contents than others. Consider indicator signals. They enable us to infer attitude contents in a way that's not available for structural representations. For instance, a predator detector mechanism can directly afford the attitude that there's a predator there. But that doesn't mean there's a 1-1 relationship between indicators and attitudes, nor that indicators always yield attitude contents directly. There can be attitudes that do not involve indicators, and others involving both indicators and representations. While detectors can point out that there's something around, structures can tell the system what that something is, what it's doing or even what it's likely to do next. Structures and indicator signals can add-up in many ways and give rise to attitudes such as "there's a tiger moving fast there".

Furthermore, a single attitude can involve multiple representations. An attitude such as "the table is red" might be achieved with the help of a |table| → *current_foveated_object* application alongside a |red| → *color_of_the_current_foveated_object* application. Other kinds of mechanisms can also participate in this pool and work as filters for the contents afforded by both representations and indicators. There can be attention mechanism pointed towards a |red-table| → *current_foveated_object* application. Such mechanism can exploit |red-table| and "abstract away" the redness of the application just like a similar mechanism could make salient the redness of a table in the world. The same goes for affections. An attitude such as "this is my father's favorite move in chess" might involve a structural representation of the board, another structural representation of the move, an attentional mechanism used to center on some aspect of the articulation of both representations and an affective element pointing towards my father. Indeed, there is evidence that fearful objects (e.g. spiders) can distort our perceptual experience, which presents them closer than they really are (Carvalho, 2021). All of these can bear on the determination of attitude contents. How exactly they add up in order to constitute an attitude is the business of a theory of attitudes, which is beyond our scope. What really matters for our current discussion is to see 1) that representations, targets and attitudes distinct *explananda*, and 2) that whatever our favorite theory of attitude content is, it need not worry about misrepresentation, for this is already accounted for at the application level.

Fodor's dictum is reminiscent of a time in which attitude contents were cognitive science's *explanans*. The contents attributed to mind states were regarded as a theoretical posit.

The task of explaining how those states could bear any content became a foundational task that philosophers took for themselves. However, cognitivism conflated issues about representational states and attitude states. To be fair, there were good reasons for adopting that stance. For instance, mental states have intentionality, and representations are a way to account for that. A successful theory of representational contents would thus allow for huge purchases. The difficulty in providing this theory is part of what lead some to reject representations (Van Gelder, 1995; Varela; Rosch; Thompson, 1991). In their view, we're better off without them, and we must prefer accounts of attitude contents in which they have no place. This is a way to reject the dictum, but it's not the only way. We don't need to fully reject any role for representations in constituting attitudes. What's needed is to reject the idea that representations should bear all the explanatory burden.[20].

By rejecting Fodor's dictum, cognitive science's *explanans* is revealed as another *explananda*. The contents of attitudes (propositional or not) are not something that scientists will get from philosophers working out foundational issues. Instead, it's something that must be empirically figured out from the pool of cognitive resources afforded by some framework. This position is evidently related to a contemporary debate regarding the nature of belief. I won't dig much into that issue, for this would take us too far afield, but it's instructive to see what would be the place of RCP in it.

Schwitzgebel's dispositionalism (Schwitzgebel, 2002, 2013) says that to believe in p is *"...nothing more than to match to an appropriate degree and in appropriate respects the dispositional stereotype for believing that p."* (Schwitzgebel, 2002, p. 253). The contrast with Fodor's dictum is stark. Most of Schwitzgebel's arguments directly address the attempt to ground propositional attitudes in language-like representational contents. Despite the similar gloss, it'd be misleading to take dispositionalism as a revival of behaviorism, for they don't share the same set of commitments. For instance, one need not reject *qualia* in order to be a dispositionalist. However, there are similarities. Dispositionalism's core idea is that implementation details just don't matter, and at least attitude-wise, they may be treated like a black box. So, whatever turns out to be true about the inner workings of human cognitive machinery, it won't change that in virtue of what beliefs are attributed. This opens the way for an all too-easy reply, which is simply to deny any real conflict with RCP (as well as any other framework that rejects Fodor's *dictum*). What grounds representational contents and attitude contents (in this case, beliefs) are disjoint. Thus, as far as determining attitude content goes, representational contents are nothing but implementation detail. In a nutshell, while Fodor's dictum puts all the explanatory burden of attitude contents in representations, the dispositionalist puts none.

But I see no reason to be commited to either position. While it's true that RCP rejects any role for language-like representations, there is a good reason for that: we must avoid FP. We need not reject any role whatsoever to structural representations in the constitution of attitudes, though. Again, this is an empirical matter. It is at best premature to start with the

[20] That's the middle way advanced by theorists like Andy Clark, among others (Clark, 1997a, 1997b)

assumption that, whatever it is we find out about the mind's reliance on structural representations, it will have absolutely no effect on attitude content determination. The philosopher's task, as I see it, is to formulate a flexible and rich framework of cognitive resources, and then allow scientific endeavors to find out how exactly these add up to become attitudes. Attitude contents are cognitive science's *explananda*. That's why one must keep in mind that the remainder of this work does not rely on any specific account of attitude contents.

An important outcome of this discussion is that, whatever representational contents buy us, it must be accounted for independently of our favorite theory of attitude contents. After all, representationalists want to use the former as a tool to explain the latter, not the other way around. Fodor's dictum made this crucial point somewhat hard to see, for it pretty much conflates issues regarding attitudes and issues regarding representations. As we've just discussed, avoiding the identity of attitude and representational contents may lead to a drastic increase in complexity. For instance, the relationship between attitude productivity and representational productivity may not be so direct as one could expect. Representational productivity is certainly an essential aspect of attitude productivity, but it may not be the full story, for attitude constitution may exploit other resources, specially when linguistic abilities are involved (Perini-Santos, 2017). Therefore, if representations refuse to yield attitude contents for free, they must be able to earn their keep in some other way. What is it that representations buy us, exactly? That question will occupy us throughout the next section.

3.4 What representations buy us

What do we earn by the insistence in making room for representational contents? Nowadays, whenever this question comes up, the background is usually a battle of frameworks: if representations are real and have something to offer explanation-wise, we should remain commited to representationalism. But if that's not the case, then we should fully embrace non-computational and non-representational approaches, such as enactivism or ecological psychology. There's a third possibility, though. The existence and exploitation of structural representations is fully compatible with 4E cognition, both online and offline. The real issue is not in making use of representational assets. It's in the classic representationalist claim that every cognitive performance involves representations and in the (now equally classic) antirrepresentationalist claim that none of it does. Both representationalism and antirrepresentaionalism are problematic, if we understand them as theses about the essence of cognition.

In what follows, I'll try to put this battle aside. Rather than arguing for a framework over the other, I'll try to show that they can share a larger set of assumptions. In particular, I'll argue that structural representations are fully compatible with the core assumptions of 4E cognition. This is not really news. Some non-classic cognitivists such as Piccinini (2021, 2022) and Clark (1997a, 2016) are trying to show this for a while.[21] Since it's harder for the

[21] I won't deal with the classic cognitivists here. They're probably not happy with my disregard for linguistic contents anyway.

non-classical cognitivist to have issues with 4E cognition, the reasoning focus on the antirrepresentationalist's resistance in accepting representational assets in a 4E framework.

The typical source of resistance comes from thinking that, by allowing the entrance of cognitive assets born in classic cognitivism, one is making room for the sub-repticious adoption of unwanted methodological principles. A concrete example of how this is manifest is in the relation between perception, action and cognition. 4E cognition takes the system-environment relationship to be the basic unit of analysis. It follows that it must reject what Hurley (2001) dubbed the "sandwich" model of the cognitive mind. Cognition is not to be taken as a mirror of the world placed between perceptual inputs and behavioral outputs. Rather, it's something that emerges out of the continuous and fluid interaction between action and perception. By bringing some (or any) notion of representation back to the picture, wouldn't this amount to a rehabilitation of the sandwich model? Cummins' distinction between targets and representations allows us to see clearly that the answer is no. The sandwich model does not follow from the existence of representational contents in the system. Rather, it amounts to a very peculiar way to describe how the system is able to fix its targets and exploit its set of available resources. How a system gets to fix and exploit its representational targets is the business of a theory of targets, not a theory of representational contents.

Representations, alongside computations, detectors, environmental couplings and the like, should always be regarded as a pool of available cognitive assets. To determine the assets employed in a given cognitive capacity or performance is an empirical matter, and this has to be done for each and every mechanism or capacity being studied. As we've seen in the examples from the previous sessions, there are cases in which both representational and non-representational explanations are available. Representations cannot earn their place because of framework-wise preferences of the researchers. They do so they participate in the best available explanation.

But there's one additional source of antirrepresentational resistance. It is the claim that representations cannot bear any relevant explanatory role. According to this view, any representational gloss is only in the eye of the beholder. We've met this reasoning before: it's the same one behind the claim that semantic properties have no causal powers of their own. This is an issue for theories based on correlation. But is it a problem for the structural semantics adopted by RCP as well? I don't think so, but that's not the reasoning I'll pursuit. Rather than having causal powers of their own, representations ought to prove themselves useful by enabling some explanatory leverage over theories that reject them. In what follows, I'll try to make a case for the usefulness and non-triviality of representations in scientific explanations of cognition. In particular, I'll argue that structural representations maximize this explanatory profit for the cheapest price in terms of theoretical commitments. If successful, this approach results in a notion of representation that's much more resilient against anti-representationalist arguments.

Cummins (1996) realized that the independence of the target and content dimensions enables the purchase of a robust account of *misrepresentation*: whenever a system fixes on a target and applies a token to it, there can be a mismatch. It can aim at a target and fail to hit it. Whenever a user fixes Belo Horizonte as a target and applies a map whose content is Belo Horizonte, there is no representational error. But if it fixes Rio de Janeiro as a target and uses the same map, then there is representational error, for that means one is taking the structure of the streets of Rio de Janeiro to be just like those of Belo Horizonte. In this sense, the token's target comprises the norm against which the token's content is to be measured. Whenever the structure of the content doesn't fit the structure of the target, there's representational error, i.e. misrepresentation. Thus, to ask for the map's target is to formulate a question such as "what's that map supposed to be of?", or "what is it you expect to find there?". Instances of appropriate answers would be something like "This map is supposed to be an accurate depiction of Belo Horizonte" or "This map is supposed to allow one to walk around Rio de Janeiro". Representations can thus play relevant and non-trivial explanatory roles because they enable us to formulate empirical hypotheses that would be unavailable in their absence. This additional explanatory dimension is characterized by the distinction between representational error and exploitation (or processing) error. To begin unpacking this claim, consider another example of how non-mental representations can be exploited at the agent-level:

(10) Igor took the wrong path because his map was excessively inaccurate; and
(11) Igor took the wrong path because he used the map upside-down.

Hypothesis (10) characterizes an inaccuracy in how the information was represented to Igor. It is, in this sense, an example of *misrepresentation.* In its turn, the hypothesis (11) depicts an error in how the available information was exploited. Mechanisms malfunctioning, bad environmental couplings, mistakes in exploring the system's surroundings, as well as less than ideal conditions, are all instances of this kind of non-representational error. Misrepresentation does not rely nor depends on how aptly the information is exploited, nor on issues with the vehicle carrying it (e.g. when a physical map gets damaged). They're possible due to the direct relationship between the structure that the representation happens to be isomorphic with (its content) and the structure to which it is applied (its target). This direct relationship can unfold a whole class of empirical hypotheses: those involving articulations of the mismatch between representational content and the target of its application. The hypothesis that Igor took the wrong path due to an inaccuracy in the map is empirically distinct from the hypothesis that he exploited the map wrongly. Even if the resulting output turns out to be similar (in both cases he ended up in the wrong place), these hypotheses would bring distinct incidental effects to the table.

Moreover, we must see that the issue is not epistemic: representational targets are distinct from actual states of affairs. Misrepresentation is not about being unfaithful to actual states of affairs. Indeed, representational targets often are non-actual states we use to think about "what's going to happen if..." or "what would it be like if...". For instance, one can have the street structure of Belo Horizonte in the 1950s as a target. However, even if the target is an actual state of affairs, there's a difference between failing to hit the target and failing to be faithful to what's out there.[22]

The independence of the representational dimension is crucial. Representations and targets have distinct *explananda*. If a system can have X as a representational target, that means it has a mechanism whose function is that of representing X. That is so because attributing the function of representing X to a mechanism doesn't mean it is actually able to satisfy what's expected of it. Rather, it means that the system will handle things as if it were successful and use the mechanism's outputs to make inferences about X. Thus, misrepresentation is completely disjoint from any sort of malfunction or exploitation error. Semantic inaccuracy does not imply nor suggests functional or behavioral failure. That's why in hypothesis (10), Igor's mistake had to be characterized as due to the "excess" of inaccuracy. The point is not that a sufficiently high degree of inaccuracy will always render unfit behavior. Rather, the point is that the map's level of inaccuracy resulted in failure due to Igor's goal while exploiting the map. A map that's "excessively accurate" could produce unfit behavior as well. Igor could get lost in all that detail that's irrelevant for his goal. It is true that this would not amount to misrepresentation, for it would be a case of failure in exploiting the map. But that's enough to establish that there's no connection between being accurate and producing fit behavior.

As another example of this point, consider how we usually give up on accuracy in virtue of tractability. That's frequently advantageous. Whenever one quickly draws a very simple map, or whenever a scientist employs a simplified model of some complex domain, one's trying to save herself from the effort of handling a huge amount of details that, while increasing the accuracy, are irrelevant for her current goals. In these cases, one's relying on misrepresentation in order to achieve better performance. Nature can rely on the same strategy. Given limited resources, an overly simplified representation of the dynamics of some predator's movements produced by a sub-personal mechanism can be the key to give the prey a chance to flee. In a nutshell, we should not be misled by the "mis" in "misrepresentation" or the "error" in "representational error". Semantic inaccuracy is neutral in this respect. The results of the level of (mis)match between representational content and the current target can be good or bad. It all depends on the system's current goal.

When given descriptions like these, antirrepresentationalists are probably tempted to disregard misrepresentation as the kind of error that can be reduced to exploitation error. After all, there can be no misrepresentation if there's no representation. Thus, whatever behavioral outcome that can be explained by misrepresentation can also be explained in terms of exploitation or processing errors. But at this point, one must realize that this is an empirical matter. In the absence of something like HPC, all the antirrepresentaionalist's has left is a crude pessimism regarding the possibility of any empirical hypothesis involving represen-

[22] This point would be a lot easier to synthesize if sentencial representations were on the cards, for we could simply say "there's a distinction between falsehood and misrepresentation", but I'm trying to avoid illustrations and analogies relying on linguistic contents.

tations be true. But this is a much weaker position to be, for HPC-like arguments enables one to claim *a priori* that any empirical hypothesis involving representations is implausible. Consider again the example of rats in mazes. Newen and Vosgerau argued that the flexible and fast way how rats learn is better explain by appealing to representational states. That is, their learning trajectory fits the available data better when characterized in term of a growing improvement on the accuracy (i.e. reduction of the mismatch between content and target) of the involved representations. In order to get rid of this hypothesis, a better one must be provided, i.e. one that fits the available data even better. That's the kind of evaluation one has to do in order to determine whether some cognitive capacity or performance involves errors of the kind (10), (11) or some complex articulation of both. But if one adopts a framework that rejects the possibility of representational states and processes by principle, one's going to discard or distort a whole class of empirical hypotheses through non-empirical means. The scientific endeavor ends up improperly constrained, for it might make it look like the only empirically plausible characterization of some behavior is by positing an error in exploiting other kinds of resources, i.e. an error of the kind (11).

Perhaps surprisingly, antirrepresentationalists are not the only ones tempted to make that move and claim that errors looking like (10) are actually like (11). Many representationalists do that as well. They obviously don't do that by rejecting the existence of representations, but they might wrongly believe that this is the only plausible path towards a naturalistic account of representational content. By understanding the source of this temptation, we'll be able to appreciate the robustness of misrepresentation grounded the distinction between targets and contents.

3.4.1 What most representationalists got wrong

Classically, in trying to account for the explanatory value of representations, many representationalists stick to what Haugeland (1998b) called *change of dimension*: within representationalism, a complex causal structure is revealed as semantically assessable. In this sense, a set of computations (or more generally, the overall dynamics within a system) can be revealed as chess playing or route planning. This dimension shifting need not be direct and can be structurally complex. Consider how many changes of dimensions there can be in artificial computational systems: electronic circuitry can be revealed as handling low-level data structures (AND-gates, OR-gates and so on), which can be then revealed as handling higher-level data types (integers, float-point numbers, strings of characters) used to comprise data structures (trees, stack, queues, etc.) of the kind we find in higher level programming languages, and so on. This can go up to the point were one is entitled to describe the system as (say) playing chess or planning a route.

For any science trying to describe cognitive capacities in terms of the satisfaction of epistemic constraints (partially or completely), this is a crucial move. It allows one to provide a scientifically acceptable explanation of how can a system be sensitive to semantically defined epistemic constraints (such as the rules of a game or the rules comprising contextually-adequate behavior). The issue with the emphasis on the change of dimension is that this purchase is compatible with accounts that trivialize or deflate the explanatory role of representations in cognition (Egan, 2020). If dimension shifting is all that representations can offer, the claim that they're in the eye of the beholder is easily justified. Curiously, many representationalists fail to appreciate this point. To see how, let us start by considering what Haugeland (1998f) tells us about the explanatory reach of biological norms:

> (...) there is another important distinction that biological norms do not enable. That is the distinction between functioning properly (under the proper conditions) as an information carrier and getting things right (objective correctness or truth), or, equivalently, between malfunctioning and getting things wrong (mistaking them). Since there is no other determinant or constraint on the information carried than whatever properly functioning carriers carry, when there is no malfunction, it's as "right" as it can be. In other words, there can be no biological basis for understanding a system as functioning properly, but nevertheless misinforming (...) (Haugeland, 1998f, pp. 309–310)

Haugeland's point is that biological norms do not allow us to talk about biological organisms that are "mistaken" in the sense that they're operating adequately, with no malfunction, under ideal conditions and, nonetheless, working with inaccurate information. From this vantage point, the only way to render misrepresentation naturalistically plausible is to understand it as derivative of other kinds of non-representational errors (i.e. malfunction, less than ideal conditions, etc.). To get a glimpse of what's behind Haugeland's pessimism, we must consider how some classic theories of representations approached misrepresentation.

To begin with, let us quickly consider the case of *conceptual role semantics* (CRS). Though it is not the most-well succeed endeavor among representationalists, it renders the point I'm trying to make specially clear. As we know, CRS attributes representational content to a state or process in virtue of how that state is used within the system. In its turn, this pattern of usage is determined by the connections - sometimes called *epistemic liaisons* - that a token bears with every other token that the encompassing system is able to instantiate. In Cummins' terms, such patterns determine the set of targets that the token is going to be applicable to. Therefore, the representational content of some state is, by definition, identical to the set of targets it is applied to. This rules out the possibility of grounding misrepresentation in the mismatch between target and content.[23] In CRS, to have the content |x| is identical to have x as a target. The content of |x| is fixed by how it is used, that is, by the target against which it is applied to. Therefore, there's no such thing as "misusing" |x|, for the content of |x| is a given of the set of epistemic liaisons afforded by |x|, and this set is nothing but the set of possible uses for |x| within the system. There's simply no room for mismatch between what |x| means and whatever |x| is applied to.

[23] These remarks refer to CRS conceived as a theory of representational contents. Whether CRS is a good choice for other *explananda*, such as attitude contents, is another matter entirely.

In what sense can we say that this approach trivializes the explanatory role played by representations? To put it rather crudely, Cummins claims that CRS is a valence theory (in reference to the valence bond theory from chemistry). In his words:

> CRS thinks about content as we might think about valence. Imagine a theory that tells us what bonds with what in what proportions. We could simply list all the possibilities (assuming they are finite). Or we could do this: assign a positive or negative number to each radical, and state the following rule: any combination that adds up to zero is a compound. Valences are a kind of fiction in this theory (multiply them all by a constant and the theory remains unchanged). Specifying something's valence is simply a convenient way of specifying its bonding potential without actually having to mention all the other elements and all the proportions explicitly. (Cummins, 1992, p. 117)

In Cummins' view, CRS results in a notion of content that is a kind of fiction. To specify the content of a representation becomes nothing more than a convenient way to refer to the set of the possible *liaisons* without making the whole set explicit. In this picture, the representation's contribution can't go beyond that of glossing complex causal structures in semantic terms, i.e. dimension shifting. Thus, it is no wonder that, when it comes to characterizing psychological explanation, it feels so easy to put this semantic gloss aside (or move it from the object of study to the eye of the scientists studying it) and claim that a description of the causal dynamics is all we need. From a methodological perspective, such a gloss can surely make our lives easier, but it's not really explaining anything. All the explanatory burden is being carried by the representation's causal role, rather than its contents.

How can a friend of CRS make room for misrepresentation? Since CRS conflates the target and content dimensions by definition, one has no choice but to bring some new, non-semantic, element to the table. As Perlman (2002) noticed, the strategy boils down to using this new element to ground some distinction between episodes of representation tokening. There are the tokenings that establish representational contents, and the tokenings that apply (i.e. exploit) the already established content. Thus, while content-establishing use makes no room for representational error, content-exploiting use does, for in some cases the system may end up applying the established content to something in the world it is not an accurate depiction of. The result, however, is a rather different conception of what misrepresentation is. What we have is not a purely semantic account of misrepresentation that buy us an extra dimension for empirical theses. Rather, it is something akin to the antirrepresentationalist approach of reducing misrepresentation to other kinds of error. For instance, rather than relying on the target/content distinction, maybe we could rely on a distinction between ideal and actual use? Consider this account from Piccinini (2022):

> (...) misrepresentation occurs when a system activates a representation, targeting a stimulus, which makes worse predictions about incoming data about what a stimulus will do and how it will appear under various possible conditions than an alternate representation that is also available to the system. (Piccinini, 2022, p. 11)

Piccinini is definitely not a champion of CRS.[24] But his account is an instance of the kind of move that friends of CRS could use in order to accommodate representational error even in the absence of the content/target distinction. He finds a spot in the fitness of the system's processing. Piccinini uses the system's own capacities as source of the norms against which to measure the accuracy of its representations. He thus grounds a kind of competence/performance distinction. But rather than relying on some abstract ideal of competence, the norm relies on the system's own bounded capacities: if it tokens |x| at a place where it would do better by tokening |y|, then it is misrepresenting |y| as |x|. Unfortunately, this approach implies that whenever the system applies |x| to a target, if |x| is the best it can do, then |x| is accurate. It is regarded as an accurate representation even if |x| is not an accurate depiction of its target. In other words: even if |x| and x may mismatch if compared independently of the system uses - this fact bears no explanatory weight in accounting for the system's behavior. Thus, there's no room for explanations such as "the system did its best and its fastest *because* it misrepresented its target as simpler than it really is in that particular occasion", e.g. because it successfully traded accuracy for speed. That's why in Piccinini's account - just like in CRS or in non-representational approaches - every instance of representational error boils down to exploitation error or reasoning error. In a nutshell: if the system is at its best, by definition, it is not misrepresenting. But if the system is not at its best, the error is not representational. We can't even formulate the hypothesis that the system is not at its best because it misrepresented something, for what counts as episodes of misrepresentation rely on less-than-ideal behavioral or functional performance. In Cummins' terms, it defines semantic accuracy in terms of behavioral or functional effectiveness: a correct representation is a representation that renders effective behavior (1996, p. 27). By conflating semantic accuracy and effectiveness, we rule out the possibility of effective inaccuracies (misrepresentations that enable the system to achieve optimal performance) or costly accuracy levels (situations in which speed and tractability should be a priority).

Another concrete example of this strategy can be found in the famous teleological approach of Millikan (1987). In trying to make room for misrepresentation, she brings adaptability to the table. Roughly, what fixes representational content are episodes of tokening that resulted in adaptive behavior to the species' evolutionary history. They become the norm against which other tokenings can be measured. If tokenings of R resulted in adaptive behavior, then R is the token's representational content. Once this is established, whenever the system tokens R in order to hit something other than R, that amounts to misrepresentation. The CRS-like doctrine that representation is a function of how the state is exploited within the system remains.

We can see a rather similar move in Millikan's approach because what grounds semantic accuracy is roughly adaptive behavior. In other words, semantic accuracy derives from behavioral fitness. From this vantage point, Haugeland's reasoning is flawless: biology makes

[24] I won't dig into Piccinini's account because much of what I say here is compatible with his work.

no room for the possibility of misrepresentation, unless it is grounded in non-semantic errors. Not only this breaks the independence among the semantic and functional dimensions, but it also presents a serious issue for the representationalist endeavor. The representationalist's goal has always been to explain functional and behavioral performances with the help of representational states. But in this picture, the explanatory order is upside-down, for we are attributing contents by relying on functional and behavioral performances. It is no surprise that antirrepresentationalists feel comfortable enough to ask: if functional and behavioral mistakes are already accounted for, why should we add a semantic gloss at all? It adds nothing to the explanation, after all.

That's why we can safely claim: it is the absence of a purely semantic account of misrepresentation that opens up the possibility of questioning the explanatory role of representations. Consequently, in order to resist the antirrepresentationalist claim, one must show that the representational dimension can be established independently, and that's exactly what Cummins' distinction between target and content buys us. Given the independent way how representational contents and targets are established, it follows that they can mismatch. The pressure that could lead us into grounding semantic error in functional or behavioral errors has no chance to come up, for there is no need to forge any gap between content-fixing and content-exploitation usage. *Pace* Haugeland, biological norms do accommodate the kind of error in which a mechanism can work under ideal conditions, without malfunction and nonetheless work with wrong information.

In this picture, despite the examples provided, it is still empirically possible that all empirical theses involving representations turn out to be false. However, once there is no HPC preempting the existence of representations playing non-trivial explanatory role, one cannot really reject them *a priori*, for their presence and effects are now a purely empirical matter. Whenever they participate in the best available explanation for some cognitive performance or capacity, there will be no reason to avoid them. This shows that there is a non-trivial place for structural representations in cognitive science. We can now do the same with representational redescription. Is there any reason to think that something like that actually happens?

3.5 Apes, humans and representational redescription

We've seen some examples of plausible candidates for real world representational accounts of cognitive performances and capacities. Time to do the same for representational redescription as presented in the previous chapter. Is there reason to think that redescription processes actually take place in human cognition? The hypothesis earns its keep for its potential role in elucidating what distinguishes human cognition from that of non-human animals such as apes. As with representational contents, the fact that representational redescription has self-sustained plausibility is important for the purposes of this work. If the entire set of tools with which RPC works were required solely to deal with RP, my credit card limit would need to be much higher than it is.

In what follows, I'll suggest that representational redescription — as well as encoding, task-embedding and the like — enable the formulation of more refined and (hopefully) fruitful empirical hypotheses regarding the continuities and differences of human and non-human cognition. As discussed in chapter 1, we seem to be a very peculiar kind of creature, and we're yet to figure out what exactly such peculiarity consists of. We exhibit intelligence in both online and offline cognitive performances. Whether we're trying to figure out a better way of using a hammer or planning a vacation, our peculiar commonsense and situation holism is frequently there in the background. But alongside the depiction of what's peculiar to human cognition, comes the question about the nature of this gap. What exactly underlies it? Is it something innate, developed or both?

Penn; Holyoak; Povinelli (2008a) draw a useful picture of the debate's contemporary form. It comprises two distinct dimensions: first, an empirical characterization of the differences between human and non-human cognition. Second, a characterization of what underlies such differences. In order to assess the latter, we need to first understand their view of the former. While presenting the first dimension, Penn, Holyoak and Povinelli (PHP for short) concentrate on the idea that human cognition is distinctive because it goes beyond observable features. This difference is manifest in lots of distinct domains: causality, sameness/difference recognition, higher order spatial relations and so on. As an example, here's their characterization of the difference in how human and non-human animals recognize sameness and difference:

> Humans not only recognize when two physical stimuli are perceptually similar, they can also recognize that two ideas, two mental states, two grammatical constructions, or two causal-logical relations are similar as well. (Penn; Holyoak; Povinelli, 2008a, p. 111)

Roughly, the idea is that we can make such judgments relying on information not directly available to perception. We can make use of "unobservable" features and attribute them specific roles such as that of being a cause. That's what we do whenever we claim that an object fell and broke because of its weight. The point is not to deny that apes and other non-human animals can't handle or even make instrumental use of the effort required to lift some object. Rather, the crucial remark is that humans can regard weight as an intrinsic property of any object, detaching its understanding from mere effort-to-lift measurements. Povinelli (2012) provides extensive empirical evidence that apes are unable to handle weight in this fashion. In order to summarize these differences, he formulated the following table:

Table 2 – How the human body and brain "represents" weight — Adapted from Povinelli (2012).

1	Anti-gravity muscles	The mass of many muscles in the body quickly atrophy in response to weightlessness

2	Effort required to lift object	The sensory-motor system tracks the force used to lift a given object
3	Effort experienced in lifting object	Humans (and other animals) have phenomenological experiences of effort lifting and moving objects[25]
4	Feed forward model of effort required to lift object	The sensory-motor system represents the predicted effort needed to lift a given object
5	Causal representation of a relation involving "weight"	The perceptual system represents the causal connections between object and outcome (organisms may use size, weight, shape, effort sensation, or any cluster of these to learn to predict an "effect" from a "cause")
6	Effects of object weight across contexts, or `f(WEIGHT)`	The human mind represents "weight" as a causal mechanism common across perceptually distinct categories
7	Linguistic representation	Humans generate linguistic (symbolic) labels for the causal mechanism of weight ("weight", "heavy", "light", etc.)
8	Metaphorical expression	Humans generate linguistic (symbolic) descriptions that connect both the phenomenological aspects of weight and `f(WEIGHT)` to nonliteral contexts

Povinelli draws a clear line between human and non-human cognition at row 6. Humans and apes share the cognitive capacities 1-5, but only humans have the capacities 6, 7 and 8. Povinelli's work concentrates on capacity 6: only the human apparatus seems able to handle things like cause, weight or color across perceptually distinct categories. He uses the notation `f(x)` to emphasize this point. Thus, being able to handle some domain-specific relation involving weight through perceptual cues is different from being able to handle `f(weight)` in multiple domains.

Despite being clear about this, the table is somewhat puzzling at first. To begin with, not every element of the table is a plausible candidate for mental representation, and Povinelli is aware of at least one such case. He knows that the first row ("anti-gravity muscles") depicts nothing recognizably representational.[26] Moreover, Povinelli's use of the word "representation" is vague and not only seem to conflate the target and content senses, but also encompasses cases that, within RPC, could be accounted for indicator signals or other kinds of cognitive tools. Tracking the effort required to lift and object (row 3) is a clear example. Thus, we must be careful when interpreting the table. Unlike what we've been doing so far, Povinelli

[25] This is clearly unrelated to accounts of the system's underlying cognitive apparatus. I kept it only to be faithful to the extent of Povinelli's work.

[26] He says: *"Just for fun, I've started with something that isn't really a 'mental representation' at all: the so-called 'anti-gravity muscles' of the human body."* (2012, p. 10). Not sure what's fun about that, though.

was not trying to find out whether the listed capacities actually rely on some non-trivial sense of representation. Rather, he takes the word in the fuzzy sense that we sometimes find in classic literature: to entertain a belief x is to represent x (which conflates attitude and representational contents), or, in more basic cases, whenever there is some causal intermediary that helps to handle x (e.g. an indicator signal), the mechanism is also described as a representation of x. We need thus the proper terminological adjustment, and that can be done by taking the table to be about targets: there is evidence that apes can target certain domain-specific causal relationships (row 5), but there is no evidence that they can target weight in a more abstract sense across contexts (row 6). This way, we do justice to Povinelli's point without giving up on everything we've discussed so far regarding what's (possibly) representational and what's not.

Despite these small shortcomings, and given the necessary terminological adaptation, the list is useful, for it illustrates the empirical possibility that these states can be produced concomitantly within the same system. In other words, the fact that a system becomes able to harbor a causal representation involving weight (row 5) doesn't mean that its sensory-motor system can't harbor another kind of representation (row 4) used to predict the effort needed to lift a given object. This is consistent with the claim presented in chapter 2: the fact that new abilities (or schemes) can result from the articulation of simpler abilities does not mean that the comprising assets are lost. But how exactly are we to understand the relationship between what's described in each row? The table depicts a route towards increasingly "higher", abstract and domain-independent cognitive capacities. This will lead PHP to hypothesize that human and non-human cognition can be graphed in a continuum towards a *physical symbol system* (PSS) (Newell, 1976), which is essentially a kind of LOT-powered system. Their point, to be fair, is not that the underlying cognitive machinery is necessarily that of a PSS. They're not commited to any full-blooded LOT-powered system. But they regard as crucial the claim that human cognitive machinery resembles or approximates PSS's features. We're talking about features such as the capacity to deal with linguistic compositionality and systematicity, as well as distinguishing between types and tokens. Their point is that the path to human-like cognition is described as a symbol approximation. They call it the *symbol approximation hypothesis* (SAH).

As an example of what motivates PHP towards SAH, consider the capacity to handle weight as a causal mechanism across distinct domains (as per row 6). At first glance, symbolic systems handling absolute contents seem to be more suitable candidates than a system that can only rely on relative (structural) contents, as suggested by RCP. Given a suitable semantics, a symbol can refer to anything: causality, weight, triangularity, leftness. In their turn, structures are unable to represent any such universal or category. They can represent a particular causal relationship, but they cannot picture causality (or `f(cause)`as PHP would put it). In the same vein, they can represent triangles or left hands, but they can't represent triangularity nor leftness. As Cummins would put it, *"no picture can show what a left hand and left shoe have in common"* (Cummins, 1996, p. 106). Here we have a potential tension between RCP and SAH.

RCP emphatically rejects any role for absolute representational contents in mind mechanisms. But the road described by SAH seems to lead us towards that kind of content. Before moving on, we need to sort that out.

3.5.1 What's wrong with the symbol approximation hypothesis

By depicting human-like capacities as PSS-like, SAH leads us to believe that we need two distinct representational systems: a pattern completion one (as the ones we find in neural models) and a symbolic one. A development of this idea can be found in Gary Marcus. He advocates the need for hybrid models in which we find both rule-based symbol crunching and connectionist-like associations (1998; 2003). Marcus is deeply pessimistic about non-symbolic machinery being able to render out-of-distribution generalizations for even the simplest linguistic or algebraic rules. To see what he has in mind, say we train a model on sentences with the form *an x is an x* for a finite set of x's instances (person, dog, gadget, etc.). Whenever we ask the model to complete a sentence like "a dog is an...", the model is able to output "...dog". But has it really captured the underlying rule `f(x) = x`? If the model can't consistently behave according to the same rule with novel words (i.e. words that were absent from the training *corpus*), then the answer is negative. Marcus' findings suggest that, though neural models can resemble `f(x) = x` for some (increasingly larger sets of) values of `x`, none can consistently handle novel values, which means the model is closer to a huge look-up table than a rule-follower.

As Marcus thinks that neural models can't fully grasp rules like these, he hypothesizes that we need innate symbolic machinery in order to compensate for these limitations. I am inclined to agree with Marcus about the current limits of connectionist approaches (or *deep learning*, as some use to call them today).[27] Though some generalization is always achievable, consistently grasping and applying algebraic rules for not originally present in the corpus seems to be still out of reach. However, I don't think this criticism bears the same weight in discussions about the human neurocognitive apparatus. First, because as previously presented, I don't think there are symbols in the mind, let alone anything like an innate symbolic machinery. Second, I think Marcus's conclusion does not follow from the empirical evidence he presents, for he does not exhaust all the options. He claims that his results are robust (they're invariant in different architectures, with different training corpus, and so on), but, as Clark; Thornton (1997) notices, it might simply be that the ability in question depends on information absent from the *corpus*. The right information might lead to the right cognition of the given domain, be it linguistic or algebraic. Indeed, human learning may involve a lot of assets and processes with no parallel in machine learning approaches (Dehaene, 2021). Consequently, no matter how robust are Marcus' results - and how damaging they are for contemporary AI efforts — they're ill-suited to make strong claims about what human-cognitive machinery must

[27] I've made use of some of his arguments in Barth (2018) and Barth (2021). See also Marcus; Davis (2019), Marcus (2020) e Marcus; Davis (2020).

look like.

Even if this weren't the case, there are deeper reasons to question the path pointed out by SAH. There's nothing necessarily symbolic about the human capacities listed by PHP. In other words, the idea that they're better (or exclusively) accounted for by symbols or the associated classic computational models is precisely what's at stake. Thus, describing a capacity as "PSS-like" or "near-PSS" implies assuming what they were trying to establish. Framework neutral descriptions of human-like cognitive capacities should not be taken to imply or suggest the existence of symbolic machinery so directly. It is true that PHP could try to deny this implication and claim that SAH is just a way to describe peculiarly human capacities, whatever the underlying machinery turn out to be. But if that's really the case, then one better drop any mention to "PSS" or "SAH". It is not doing any work that a framework-neutral depiction of those human-like capacities couldn't do by itself. In this sense, I think PHP fell for the constraint we've talked about at the very beginning of this work: the difficulty in formulating framework-neutral descriptions of cognitive capacities.

Perhaps an alternative to try to save the symbol talking is by offloading them to the environment. As Barrett (2008) suggests PHP seem to take a broadly internal stance towards mind mechanisms. In her view, the authors ignore the possibility that whatever accounts for the SAH character of cognition can be out of our heads. Rather than relying on PSS-like machinery, we employ external symbols as tools. But PHP correctly resist the more radical versions of this idea (Penn; Holyoak; Povinelli, 2008b). What grounds their concurrence is that the difference between the cognition of human and other primates cannot be neither solely nor mostly out of our heads. Simply offloading symbols to the world would have thrown no light on that difference. As we'll see shortly, I think external symbols do play an important developmental role in accounting for how we handle the world, but not the whole story. Furthermore, fully offloading contents preempts the explanatory leverage that mental representations buys us, for it simply ignores the empirical possibility that mental representations can provide the best available explanation for certain cognitive capacities and performances.

Once we get rid of the PSS talking, we become open to a different diagnostic and a different possibility. The crucial idea is that what PHP regard as `f(cause)` or `f(weight)` cannot be merely a symbol. Rather, it is a *concept*. Let's unpack this claim, right after a quick remark: I am aware of the large and old dispute in cognitive science regarding what a concept is. However, I have neither the intention nor the need to dig into that issue. Thus, with the likely exception of those who follow Fodor (2004; 1998) into thinking that concepts have no structure nor any kind of epistemic condition, I assume that none of what follows will be polemic.[28] Even eliminativists such as Machery (2009) should not be alarmed, for as we'll see, what matters for my purposes is just the claim that we can't send a symbol to do a concept's job. What I regard here as a concept can be conceived as a bunch of cognitive assets that are somehow connected by sharing a topic. Attitudes, representations, indicator signals, know-

[28] Well, maybe those advocating a linguistic notion of concept, such as McDowell (1994) would also be worried about this. But they probably wouldn't bother reading about philosophy of cognitive science anyway.

how and whatnot. We also need not worry about how such assets are structured or articulated. Thus, whether they comprise necessary and sufficient conditions, prototypes or theories will bear no weight on the discussion.

Now back to the main point. Consider what it takes to have the concept DOG. It may involve representations of what a dog looks like, attitudes about what dogs like to eat, knowing-how to walk them, and so on. Something similar goes for the concept of LEFTNESS. It may involve memories of a time in which one had to look at her hand in order to determine which one is the left. Perhaps with the help of some habit (writing), accessory or birth mark. This can grow into the knowledge that her pet also has left legs, that the car has a left side and so on. The concept of LEFTNESS may comprise all of this cognitive knowledge. None of it requires symbols. It surely requires the capacity to articulate cognitive resources in certain ways, but the lack of symbolic assets brings no principled issue.

The same goes for concepts like CAUSE and WEIGHT as they appear at row 6 of Povinelli's table. If we think about CAUSE as |cause|, that is, if we think about concepts as representational content, then we'll be lead quickly to the idea that this representational content must be symbolic and to believe that there's some radically different representation machinery at work in human cognition. But that's not necessary. What's distinctive of human cognition and ape cognition may comprise lots of differences, big and small. We may be able to have representational targets that apes don't. We may be able to represent those targets using schemes that apes don't have. We may even share representational content with apes, but maybe we exploit them with consumers that are sensitive to structural aspects that their consumers are not. Moreover, we may be able to articulate distinctive kinds of cognitive assets in distinct ways by, e.g. encoding or task-embedding. There can even be a human way to cook up attitude contents from representations, indicator signals and other assets.

This is not a speculation about which of these is true, if any. The matter is obviously empirical and won't be settled by armchair thinking. Rather, the point is to show how the continuum towards "symbol-like" machinery can be misleading. Rather, the distinction between human and non-human cognition may be along the dimensions unfolded in the previous chapters. For instance, given any domain, it is possible that we share representational capacities without fully sharing the set of available exploitative capacities. We may share lots of representational targets and yet handle them in very distinct ways. This enables us to talk about discontinuity in human cognition in a more refined way. We can do stuff that non-human animals can't, but that need not imply any big jump in the kind of mechanisms we employ. In other words, we can render both human and non-human cognition discontinuous in the sense that they don't rely on the same tool set, but continuous in the sense that such tool sets are much more alike than we're used to thinking about them when framing the issue through the lens of a specific framework.

At this point, it seems like we have a plausible alternative to SAH. We could maybe stop the reasoning here with a claim like this: it is an open question whether what PHP calls f(weight) is actually a symbol for weight or a concept of WEIGHT.[29] But we can't stop here, for the very good reason that it is not true. Regarding f(weight) as a concept is the only plausible way, or so I intend to argue. We have to take another step and show that walking towards symbolic systems, *pace* PHP, won't buy us what's needed. In other words: symbol systems are not good rulers of what's distinctive about human cognition. This is not a point against PSSs *per se*, nor a point against the possibility of modeling cognitive capacities with symbols (though there are limits, as FP shows). Rather, the point is that even if we find out that humans could have a fully-fledged symbolic machinery that somehow handles FP, the question about what underlies the difference between human and non-human cognition would remain intact.

PHP seem to assume that the capacity of thinking about f(cause) is in some sense associated to the capacity to forge a symbol referring to f(cause). A symbol for weight or cause is no different from a symbol for dog or cow, though. It refers to the thing without actually comprising any information about it. But the explanatory relevance of something like f(cause) or f(weight)is not in the possibility of having symbols referring to it. Rather, it is in what the system knows about them. Having a symbol pointing to CAUSE is like having a symbol pointing to COW. The symbol itself tells you nothing about neither causes nor cows. It won't help the system to recognize them nor guide it while thinking about them. Thus, even on symbolic systems, what accounts for the system's capacity of handling f(cause) (or f(cow), or f(weight)...), is not the symbol, but whatever the system knows about its referent, i.e. its concept.

In a purely symbolic system, such cognitive knowledge would comprise lots and lots of sentence-like absolute contents stored through symbolic medium. A CAUSE in a symbolic system would be akin to a CAUSE in a non-symbolic system, but seriously constrained, for all the knowledge comprising CAUSE would have to be expressed using only sentence-like formalisms. That's exactly what McCarthy and Hayes were trying to achieve when they first faced FP, by the way. Again, I'm aware that PHP are not advocating for purely symbolic systems. Their view is probably best described as a hybrid one, in which some sort of symbol system lives alongside a non-symbolic one (as per Marcus' proposal). But if symbols can't buy us CAUSE, that means they can't buy us f(cause), for these are the same. Why should we bring symbols to the discussion at all?

In general, whenever PHP talk about "PSS-like" systems, what they have in mind seem to be systems capable of doing logical reasoning. And symbolic systems do seem to buy us that. They allow for language-like compositionality and language-like systematicity. We could then, at least in principle, forge a truth-conditional semantics that allows us to put complex articulations of symbols in pair with complex articulations of words in natural language. The

[29] Some might feel tempted to reject this distinction by saying that PHP's hypothesis is able to accommodate it. That is, regarding f(weight) as WEIGHT is actually a way to approximate a PSS. First, it is not, and we're about to see why. Second, if being "PSS-like" goes for this, it probably goes for pretty much anything remotely similar to human cognition. This means that the talking about PSS is not really throwing any light and should be dropped.

problem is that, while such semantics may tell us how symbols can be managed in truth-preserving ways, it won't do the same for concepts. When you deny the symbol for dog, you get a pointer towards the set of all things that are not dogs. But if you deny a concept DOG, what you get is not a concept of NON-DOG, i.e. information about all the things that are not dogs. In the same vein, no truth-conditional semantics will tell us what results from the conjunction of GOOD and DOG. In order to grasp that, one must have some information about what GOOD and DOG are. Remarkably, this has the potential to weaken arguments like that of Marcus' even more, for the capacity to fully grasp `f(x) = x` won't buy us what's needed.

In a nutshell, the problem with the reliance on symbolic machinery — even by means of approximation — as a way to express what's peculiarly human in cognition, is that, at the end of the day, PSSs are silent about what really matters. They don't tell us how to merge concepts, nor can they throw light on what exactly comprises the underlying difference between human-WEIGHT and ape-WEIGHT. With this in mind, my suggestion is that we'll do better getting rid of the symbolic talk and concentrating on more framework-neutral depictions of the differences in overall cognitive capacities.

3.5.2 The relational redescription hypothesis

Despite the taste for PSS-like machinery, the way PHP frame the question of the gap between human and non-human cognition is distinctive. Rather than contrasting linguistic and non-linguistic capacities, they concentrate on the long gradient of possibilities within. We already had a hint of how the ideas described in chapter 2 can throw some light on what underlies some important differences located along this gradient. Ape's cognitive machinery, inasmuch as it exploits mental representations, might use distinct schemes, distinctive tokens or distinct exploitation strategies. More importantly, these ideas comprise a common framework in which different perspectives can forge different empirical hypothesis about what makes us human or why apes can't handle language. One need not regard humans as smart and apes as dumb. We can say that humans and apes can do a lot of smart things without giving up on the claim that humans do it in a way that allow them to go further. In other words, similar behavior need not imply similar cognitive apparatus. We have thus a framework in which the minutia of Povinelli's relational redescription hypothesis (2012; 2000) can be articulated.

Povinelli presents his understanding of what relational redescription is, as well as his arguments using as example the evolution of social cognition. For many, social cognition must be grounded in the so called *theory of mind* (TOM), which requires peculiarly human abilities, such as the capacity to estimate what other agents are thinking and what are their beliefs or intentions. Povinelli claims that arguments for TOM usually rely on analogies: we see non-human animals exhibiting human-like behavior and posit the presence of human-like cognitive machinery in order to explain it. This idea is not new. Povinelli traces it back at least to Hume, and take it to be entertained by figures like Darwin. As an example of argument relying on this kind of analogy, we can consider the evidence of deceptive behavior in apes. This can be found in Waal (1992) and Waal (2007). The cases presented by de Waal are sometimes regarded as strong evidence for the view that one cannot deceive someone else unless one can somehow reason about that someone else's mental states.

In stark contrast, PHP thinks that social cognition can be articulated without TOM. Here's how Povinelli presents this view:

> (...) in our model, the behavioral forms that primatologists are fond of calling deception, empathy, grudging, and even reconciliation, all evolved and were in full operation long before there were any organisms that could interpret these behaviors in mentalistic terms. In short, these behaviors did not evolve because our earliest mammalian and primate ancestors possessed the means of representing the minds of their fellow group mates. Rather, these behaviors evolved because they became inevitable as selection honed psychological–behavioral systems to maximize each group member's inclusive fitness. (...) we believe that our new psychological systems for representing and explaining already-existing behaviors in terms of mental states were woven in alongside these ancestral systems. Thus, the reinterpretation model posits that most of the basic behavioral patterns present in our ancestors (as well as the psychological mechanisms for producing them) remained undisturbed by the evolution of theory of mind. (Povinelli, 2000, p. 59)

In a nutshell, while the TOM hypothesis explain social behavior by positing the existence of human-like mental states about someone else's mental states, the relational redescription hypothesis claims that social behavior emerged without relying on anything peculiarly human. Instead, capacities such as TOM came later by and enabled humans to reinterpret their own social behavior in TOM terms.[30] This opens up the possibility that humans and apes share much of what underlies their social behavior, but only humans are able to interpret that behavior in TOM terms.

PHP's idea is not limited to social cognition. There's a more general thesis underneath it. Its crucial claim is that what we regard as "high level cognitive capacities" comprises the human capacity to reinterpret evolutionarily old abilities and mechanisms used to cognize social or causal domains in less-than-general terms. As we have seen, such reinterpretation is usually taken to consist on mechanisms able to target ancient systems and render then more general and domain-independent. That is, what was once grounded in the exploitation of observable features and relationships can now be "objectified" into an unobservable property. That's how we can tell what's similar between a left hand and a left shoe, or between a bird's nest and a dog's doghouse. This is the spot where the discussion from the previous session must be placed. The idea that some properties can be abstracted away from their domains is usually regarded by PHP as their redescription into "symbol-like" medium. However, we've seen why that's misleading. What we need are not symbols, but increasingly complex and richly articulated concepts.

This is also the spot where the extra degrees of freedom yielded by RCP pays dividends. The many possible relations between cognitive assets previously described — redescription

[30] This idea fits well with the hypothesis that TOM, as well as causal reasoning, language and similar capacities are actually a kind of cognitive technology. We'll develop this point further in the next chapter.

through structural analysis, encoding, task-embedding and the like — provides the essential tool-set that the system needs to engage in reinterpretations of both its own dynamics and the dynamics of others. In other words, it can render new understandings about itself and everybody else. While it is true that Povinelli relied on Karmiloff-Smith's original idea of representational redescription, [31] the RCP version retains the same goals, even though it rejects the idea that redescriptive processes necessarily run towards absolute contents and symbolic machinery. They can provide more abstract understandings, but they can also forge new ones, even though not necessarily of a more abstract sort. This means that, within RCP, one can go beyond formulating models that put apes and humans in a linear "vertical" trajectory towards increasingly higher abstracts levels (as table 1 suggests). One can also model "horizontal" disparities as well. What I have in mind are different takes on the same kind of phenomenon or cognitive capacities. Remember, for instance, the warning from De Waal (2016) about how the capacity to handle tools in order to solve problems can take many forms, mainly due to the creature's particular *Umwelt* This kind of difference can also be accounted for in terms of representing things under different schemes, exploiting these schemes with different strategies or perhaps there can be species that can only encode what others can represent, and so on. The space between purely behavioristic accounts of non-human animal cognition and full-blooded folk-psychological accounts can accommodate a large, multidimensional gradient of intermediate possibilities. Now, in order to make these claims a bit less impressionistic, lets draw a sketch of how a broadly non-symbolic story could go.

3.5.2.1 Domain bootstrapping

Structural schemes have target-domains. Within those domains, they're at their best, in the sense that they can fully play their explanatory power. However, most examples involve only perceptual-level domains (natural images, cognitive maps and so on). But whenever one talks about "higher" cognitive capacities, one usually means episodes of offline cognition. Thus, unless we want to seriously constraint representations' explanatory role to online settings, a large portion of the story must be about the "higher level domains". What does it take to grasp one of them?

Fortunately, structural representations are broadly compatible with both online and offline capacities. They can play the role of stand-ins at both levels. This means that we don't need to posit different kinds of cognitive resources only to account for this difference. Whether we're talking about environmental affordances or a sudden realization that there's a better way to plan the next travel, structural representations can be there to help. For instance, the same representational resources that figure in online cognition can also be used in offline simulations, in order to predict outcomes. That is, the system can easily apply lower-level knowledge about the world in order to guide offline simulations. This is useful because, whenever some aspect of the world is absent, you'd better have something else to put in its place,

[31] As stated in (Povinelli, 2012, p. 297).

and structural representations are likely to provide the most powerful tool for the cheapest price.

Something along these lines comprises Barsalou's *simulators* (Barsalou, 1999, 2003), but we can also find similar ideas in Friston; Mattout; Kilner (2011) and Clark (2016). While developing their suggestions, Barsalou, Friston and Clark have their own frameworks in mind, but all of them work with the idea that the representations used in online cognition can be repurposed and exploited by mechanisms dedicated to offline cognition. Trajectories throughout some domain state space can be thus used to think about it even in situations where the agent is causally detached from that domain. Furthermore, cognitive machinery can learn how to articulate permutations and variations of those trajectories, thus performing simulations that help the agent tell "what's gonna be like" when somethings happens. The upshot is that any representational resource employed in action and perception (i.e. online capacities) is also available to play non-trivial roles in planning, reasoning and thinking in general. Moreover, structural representations ensure that, if someone eventually figures out a complete story about how structural representations (alongside indicator signals, environmental couplings and other cognitive tools) can handle higher, offline, cognitive abilities such as those involved in mastering natural language, then that story will not suffer from FP.

However, not every capacity to negotiate a domain can be regarded as a simulation of lower level perceptual ones. Offline cognition usually involves rather peculiar domains. Consider the classic example of playing chess. The chess domain comprises very distinctive sets of rules taking place within a heavily constrained environment. The domain is out there for us to learn about just like the domain of natural images. However, we can easily understand where the structures comprising natural images came from. But chess? How exactly this kind of that domain comes to be?

We can find a clue for a possible bootstrap story in the classic work of Clark (1997a). Clark's idea relies heavily on the human capacity to create and maintain lots of external symbolic structures. The crucial claim is that external symbols (sounds, words, drawings, objects) can be regarded as discrete labels for complex concepts. Consider again concepts previously discussed such as CAUSE or WEIGHT. These concepts comprise lots of knowledge. External labels such as words ("cause", "weight") or pictures enable us to stabilize everything we know about the respective topics under a kind of indivisible unit. This unit can then be a primitive element in increasingly complex, higher-order structures. In this sense, by articulating their elements we forge new domains in which the possible permutations get stabilized. By associating a portion of knowledge to a public symbol, we can both associate those "objectified wholes" with each other (as we do in "healthy weight"), and shape our own thinking by learning to associate what were, up to this point, disconnected portions of our knowledge (in the example, those about HEALTH with those about WEIGHT). As this kind of association goes on, we get new structured domains, and these can be cognized through specific models (maybe using specific representational schemes), that capture the domain's expressive power and systematicity.

A common example of this phenomenon comes from the spatial domain: if we put washed itens in box 1 and unwashed itens in box 2, we can then think about what to do with box 1's itens without attending to their specific perceptual properties. By sticking to their presence in box 1, we're abstracting those other properties and simply taking them to be a set of washed stuff. This is the kind of capacity that may comprise concepts like WASH.

A more concrete example of same point can be found in Thompson; Oden; Boysen (1997). Thompson's experiments have shown that apes exposed to token based symbol systems can make limited, yet impressive, use of them. The apes in question could recognize and use labels for "shoe" and "cup", and they could also recognize second-order sameness relations between different distributions of those. In sets comprising, say {cup-shoe; shoe-shoe}, we have two first-order relations (one being of difference, for a cup is not a shoe, and the other of sameness, for a shoe is a shoe), and also one second-order relation, that is of difference (for the relation between a cup and a shoe is different from the relation between a shoe and another shoe). In the resulting picture, we have a cognitive engine capable of representing the structures of both first-order and second-order relations. However, when assessing this kind of evidence, PHP claim that:

> (...) these experimental protocols lack the power, even in principle, of demonstrating that a subject cognizes sameness and difference as abstract, relational concepts which are (1) independent of any particular source of stimulus control, and (2) available to serve in a variety of further higher-order inferences in a systematic fashion. A functional decomposition of the S/D and RMTS protocols reveals that the minimum cognitive capabilities necessary to pass these tests are much more modest. (Penn; Holyoak; Povinelli, 2008a, p. 112)

Indeed, what the experiment shows is not enough to establish the presence of a proto-symbolic mechanism. But that's only an issue if we stick to PSS-like view o the mechanisms involved. The apes from Thompson; Oden; Boysen (1997) did not acquire full-blooded PSS-like capacity to generalize over rules, but they did acquire the capacity to deal with certain second-order regularities in some domains. That's a possible case of representational redescription. The possibility that human-like cognitive capacities comes from the cumulative cognitive technology is not at all ruled out, and there's a reason to be optimistic about that: we don't need symbols to encode the structure of a domain. We can encode, for instance the phrase structure of the current linguistic input, and this can guide us in our thinking (that is, on how to connect different parts of our knowledge as indexed by linguistic labels), somewhat like a recipe can guide cooking activity. We don't need to suppose that recipe and the cooking are made of the same stuff for that to happen. Cognitive machinery comprised of a plurality of domain-specific representations allows us to see how we can gradually grow towards cognitive capacities increasingly complex without the need to postulate a specific proto-symbolic mechanism.

Clark's idea is that this basic phenomenon can be repeated recursively for indefinitely many times. This has the potential to result in cognitive capacities of increasingly higher level, and the structures of the emerging domains are its output. We can thus objectify clusters of properties, and relate these with others objects. From this, new structured high-ordered domains can emerge, and these can be gradually learned and represented by using the same kind of cognitive mechanism that we already have.

It is tempting to think of such "public symbols" as linguistic elements, and Clark's idea may indeed capture something necessary to account for the emergence of linguistic capacities. But his idea is not restricted to language nor intends to fully account for it. Rather, many distinct higher order domains may emerge out of processes like this. These domains may underpin social practices, institutions and of course chess. Importantly, I don't think Clark tries to deny a divide between humans and other animals. He would not deny that there is something in the human brain that allows us and (as far as we know), only us to learn a full-blooded language. What's at stake is the nature of the divide, and how much of a genetic "starter pack", comprises the relevant properties of the human brain.[32] Thus, Clark's insight can account for the way we can forge increasingly complex structures comprising various domains that can be then cognized by already existing cognitive machinery.

The story is not that the exposition to public symbol systems is enough to completely explain the capacity to deal with abstract symbols. We're not just chimps who got lucky and have found a nice set of tools. What's at stake is the nature of the divide between us and non-human animals. Such a gap might be large enough to establish very robust differences in cognitive mechanisms, but not to the point where we would have to assume some completely different kind of mechanism (like a proto-symbolic one), and all the evolutionary difficulties that come with it.

Clark's story is consistent with Hurley's picture of "islands of rationality" we used in chapter 1 in order to characterize islands of integrative capacities. Just like integrative capacities, the knowledge comprising concepts like CAUSE or WEIGHT can quickly grow as the system learns its way into novel domains and situations. How far it can go is of course a function of the creature's overall cognitive capacities. Perhaps only linguistic creatures can entertain concepts with full-fledged generality, but as the relational redescription hypothesis suggests, the capacity to hold a concept like WEIGHT is certainly not and all-or-nothing matter.[33] Thus, distinct creatures may be able to recognize a CAUSE in distinct domains, even though it is not able to recognize it in every domain it can negotiate.

Now, one of the cognitive abilities that PHP describe as exclusively human is that of making similarity judgments by attending to non-perceptual features. That amounts to, for

[32] I think the best candidate for what comprises our "genetic starter-kit" is provided by Heyes (2018a). Her work will be extensively discussed at the final chapter.

[33] It is not clear what could be the role of language in accounting for fully-fledged generality. I'm inclined to look at it as a crutch for exploiting cognitive assets. Once we learn how to grasp sentence-like structures, they become a tool we can use in order to communicate and write down which cognitive assets we want to exercise and in what way. In this sense, a sentence like "the elevator is blue" tell us that we should recruit ELEVATOR, BLUE and articulate them in a certain way. By the same token, a sentence template such as "X causes Y" can have any concept we're able to entertain as values for both X and Y. Thus, we can eventually read something like "elevators causes blue", but it doesn't mean we can make sense of it. We can always try, though. In this sense,fully-fledged generality is more of a linguistic tool than a cognitive capacity.

example, the capacity to recognize any two features as instances of CAUSE or WEIGHT even though the features can be found in perceptually distinct domains. Thus, whether one's talking about objects being pushed to the ground by something or a social invitation to a party being accepted, humans can recognize both as instances of a CAUSE. In the former, we refer to the cause of an object falling and breaking. In the latter, it is the cause why someone went to a party. This can go on indefinitely for any concept: humans can (fallibly) recognize whether any two mental states are similar, or whether any two logical relations are alike, and so on.

At this point, it might be puzzling that, at least in the cases where structural contents are all we have to compare cognitive knowledge, what exactly is being compared whenever two instances of CAUSE from distinct domains are regarded as similar? What could be structurally similar to a physical-CAUSE and a social-CAUSE? Clark's story buys us the following answer: concepts comprises the integration of cognitive knowledge under a given topic. Part of how this integration story involves forging a new domain. Learning about CAUSE (or WEIGHT) amounts to articulating knowledge from distinct domains into an emerging "domain of causes". As we integrate further knowledge into CAUSE, the domain of causes gets more complex and enables more and more permutations. Causes may vary enormously in their details, and the capacity to handle the set of possible permutations may be handled by a specialized domain in which these permutations takes place. All of this can be cognized by either encoding or representing the possible states within the recently emerged domain of causes. Such domain provides the state space in which we can compare the many structures comprising each permutation. This buy us enough to compare concepts from distinct domains without relying on perceptual cues, just like PHP requires. In the end, non-perceptual similarity judgments can boil-down to structural similarity within distinct domains. Let us unpack this a bit more.

3.5.2.2 Similarity judgments within emergent domains

To see how structures can handle similarity judgments, consider Churchland's suggestion that we can understand semantic similarity (or even identity) in terms of neural state space similarity (Churchland, 1998). This point emerged in the context of a debate with Fodor; Lepore (1991) about Churchland's space state semantics. Churchland (1989) suggests that mental contents are given by the position that an activation pattern occupies in an n-dimensional activation space. By the same token, latent (long-term) cognitive knowledge would be a product of the position in a weight space, i.e. a function of how the relevant populations of neurons are wired. However, Fodor and Lepore argued that neural models could never provide a robust criterion for inter-individual content similarity. Here's an example of how they put it:

> If the paths to a node are collectively constitutive of the identity of the node (...) then only identical networks can token nodes of the same type. Identity of networks is thus a sufficient condition for identity of content, but this sufficient condition isn't robust; it will never be satisfied in practice. (Fodor; Lepore, 1996, p. 147)

In other words, Churchland's account fall short of explaining how can two distinct systems be in the same representational state. The identity of those states is only possible assuming that the network has the same number of neurons arranged in the same fashion (i.e. has the same architecture). Accordingly, the same latent cognitive knowledge of any two systems can only be identical when the neurons are identically wired (they occupy the exact weight space). In the case of learned connections, this requires identical learning conditions with identical *stimuli.* As a consequence, in order to ascribe the same representational content to any two human agents, we would need the exact same level of activation in the exact same number of neurons (arranged in the exact same way), in both brains. Though theoretically possible, that is obviously impractical to satisfy, which is why the worry presented by Fodor and Lepore is sound. In order to avoid their criticism, Churchland needs a way to show that the same representational state can occur even in distinct neural populations with distinct wiring.

Churchland's way out involves giving up on the idea that representational states are a given of the exact position in an activation space. Instead, they can be purported as a function of *sets* of possible positions within the same activation space. Perhaps the best way to present the approach is by using the example provided by Churchland himself: say we're training an artificial neural network to recognize the facial distinctions among members of families: A, B and C. The training *corpus* comprises lots of pictures of distinct members of each family. The goal is to construct a network that gets some facial picture as input and is able to output the family it belongs to: A, B or C.[34] The resulting distribution of activation patterns is depicted in figure 12(a).

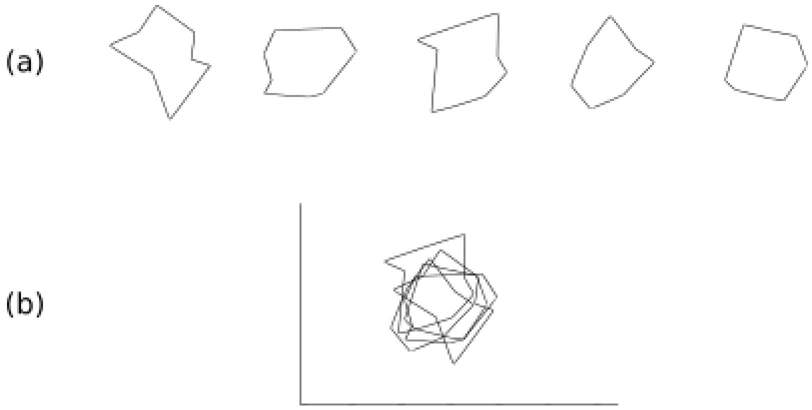

Figure 11 – Churchland's prototypical partitions.

Whenever the network is fed a picture from a member of the family A, the resulting activation pattern will be around the A region. Likewise for regions and families B and C. The weird mesh in region A of figure 12(a) is not just a partition in the activation space, but a set of them. Churchland calls such partition sets *prototypes* (Churchland, 1998). Figure 11(a)

[34] For simplicity, let's rule out the possibility of presenting the network with a facial picture of someone from a fourth family.

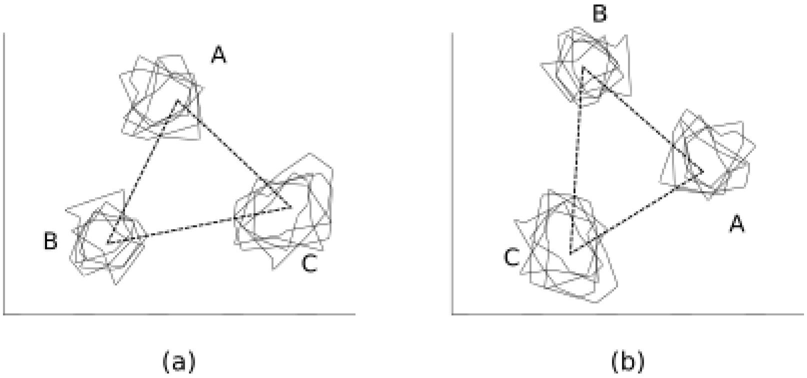

Figure 12 – Structural contents in partitions of activation spaces.

exemplifies distinct geometric partitions equivalent to distinct positions in an activation space. We can see what would be the prototypical region for the average member of a family in 11(b). [35] A prototype does not depict any specific member. Instead, it works as a center of gravity that aggregates typical features of A members.

Now say we've decided to train a second network. One with a distinct number of nodes (neurons) and perhaps even a distinct number of hidden layers. The resulting distribution of activation patterns for this second network is depicted in figure 12(b). We can see there how the prototypical regions occupy distinct portions of the network's activation space. This is the foundation for the criticism of Fodor and Lepore:how could we tell whether the two networks are representing or encoding the same thing? However, perhaps you already noticed that the triangular A-B-C structure is geometrically alike in both networks. How could that be? The answer can be found in Laakso; Cottrell (2000), who developed what is today know as *representational similarity analysis.* The technique became an important component of contemporary research in both neurosciences and AI, particularly in the Deep Learning tradition.[36]

The researchers used cluster analysis in distinct neural models to shown how similar geometric patterns of activation would appear even in heterogeneous representational spaces. Thus, the evidence suggests that the structures captured by the spatially arranged sets of prototypes are objective. In the family resemblance example, the A-B-C triangle is the (evidently oversimplified) geometric depiction of the objective structure comprising the relationship among facial cues of distinct family members. Just like the basis functions afforded by the *sparse coding* approach, the structures comprising prototypical regions are out there to be captured. In both cases, the resulting structures are candidate to further exploitation of its geometrical properties by the encompassing system. Furthermore, prototypical regions cannot be regarded as sets of indicator signals. Consider again the Palette-C scheme in figure 9: that in virtue of which a straight line becomes a human arm or leg is its geometrical relationship with other primitive elements. Likewise, what makes of B a prototypical region

[35] The positions are evidently simplified in order to be expressed as 2D drawings. Real world activation spaces are usually multi-dimensional.

[36] See for instance Abnar *et al.* (2019) and Nanda *et al.* (2022).

for members of that family just *is* its geometrical relationship with the A and C prototypes. Thus, we should not be misled into thinking that the trained networks first learned prototypes for each family and then learned their distinctions. It's quite the opposite: the prototypical regions are established as one learns the difference between the distinct families, that is, as one learns to capture the A-B-C structure in increasingly accurate ways.

The upshot is that we don't need symbols nor any machinery resembling it in order to explain how a creature can handle CAUSE or WEIGHT. We need not fully-fledged "logical and abstract" capacities in order to go beyond perceptual stimulus. Even if Penn; Holyoak; Povinelli (2008a) are right when they claim that non-human animals do not rely on non-perceptual structures in order to render sameness/difference judgments, that does not allow us to regard non-human cognition as "less symbolic" and our machinery as "more symbolic". Symbols are not helpful in characterizing the norm against which we should assess our cognitive capacities. Thus, we need not suppose the abrupt emergence of some cognitive machinery that is somehow able to render abstract and logical reasoning.

In order to allow shareable judgments of sameness and difference (i.e. judgments pointing to the same cognitive target), we need not fully-fledged linguistic promiscuity. Different systems need only to share enough of the structure domain involved in using a given concept and similar developmental trajectories are sufficient to account for that. We can thus abandon SAH and still honor all the relational redescription's goals and constraints put forth by PHP. RPC affords a story that does not rely entirely on language. One need not commit to the idea that ape minds with linguistic abilities would be nothing short of human minds. Moreover, one need not symbols in order to account for so many distinctive ways to render the world intelligible. Multiple representational architectures and distinctive exploitation techniques buy us an extra degree on freedom that our empirical hypotheses may rely on to do comparative psychology or or related scientific endeavors.

3.6 What now?

If we have succeeded in the aims of this chapter, two important purchases were made. First, we've made room for structural representations in accounts of cognition. Such representations can play non-trivial explanatory roles in a way that does not preclude 4EA's cognition most dear methodological commitments. Their non-triviality comes from the extra degree of freedom we get to formulate empirical hypotheses that would not be possible in their absence. We have a robust notion of misrepresentation and can explain behavioral outputs in terms of a purely semantic relationship between a representational content and its functional target. Moreover, the architectural traits of the employed schemes can also participate in explanations of cognitive capacities and performances, for they amount to sets of assumptions about how the world is, which means they can constrain the way a system registers the world (i.e. takes it to be).

Second, we have the possible relations between distinctive representational schemes, as well as the many possible exploitation techniques, particularly that of representational redescription. This buy us a considerably large dimension in which we can place empirical hypotheses about what's peculiar to human cognition in contrast with non-human cognition. Rather than putting all the burden under a wide (yet unique) capacity (theory of mind, language, and so on), the issue can be articulated as multiple empirical research problems. We should be allowed to ask, for instance: are the subjects learning a new representational scheme (i.e. how to deal with a new structured domain)? Or are they learning how to exploit an already-familiar domain in a new way?[37] All of what can be done while avoiding the potential problems that come with SAH. To start with the assumption that there has to be some kind of symbolic machinery in order to account for what's unique in human cognition improperly constrains the set of empirical possibilities.

We have thus achieved representational productivity without absolute contents, which means we can engage in research without the risk of raising FP. Everything we need to make an attempt at handling RP is in place.

3.7 Paragraph-by-paragraph summary

Each entry below summarizes a paragraph of the main text.

Battle of frameworks

If we want to make room for representational pluralism in cognitive science, we must show that employing multiple representational architectures is both psychologically and mechanistically plausible, in the sense that it is both elucidating and compatible with current scientific practices.

The thesis that the mind employs multiple structural schemes is what I'm calling representational cognitive pluralism (RCP), and it is able to accomodate insights from classic cognitivism and non-representational approaches such as ecological psychology, varieties of enactivism and others.

Two of classic cognitivism's core motivation for its heavy reliance on representations were 1) a conception of cognitive activity as essentially that of computing over symbols and 2) the thesis that attitudes contents were a direct given of representational contents.

A third, sometimes under-appreciated motivation is the possibility of using proprietary formats (i.e. a particular ontology) that can alleviate the tractability of complex problems: the form with which we register the world (i.e. with which we encode or represent it) matters just as much as the fact that we do it.

This is probably why classic cognitivism mitigated the importance of perception in the study of intelligence: perceptual processes were conceived as taking place before the mind's registration of the world.

[37] Notice how this is reminiscent of the old debate between piagetians and their adversaries. It is sometimes under appreciated that if a given cognitive framework does not allow this kind of empirical question to be formulated, what we have is, as Chemero (2009) remarked, a kind of hegelian move that rules out empirical questions through non-empirical means.

RCP rejects the first two mentioned commmitments of classic cognitivism, but it profits from the idea that registering the world in proprietary forms has an explanatory role to play, and to that extent it requires a theory of representational contents.

Thus, inasmuch as the third motivation is concerned, it takes issue with non-representationalists, for they insist in making no room for any conception of representation.

Likewise, iansmuch as the third motivation is concerned, it also takes issue with classic cognitivism, for they insist in a single, languagel-like way to register the world.

While the classic cognitivist worries that giving up on language-like scheme is a step back to the kind of problems faced by behaviorism, the non-representationalist worries that representations are a step back to the kind of problems faced by classic cognitivism.

Considering this, my goal is to find a middle way and sketch a picture in which structural representational contents can bear non-trivial explanatory roles without giving up on the 4EA cognition's tenets.

Thus, I have no intention to claim that every cognition is representational, only that it is empirically plausible that representations can play non-trivial explanatory roles.

Time to start paying the huge debts I incurred in this section.

Representations and targets

The first step is to be familiar with Cummins' distinction between representations and targets.

When a system uses a token T as a stand in for something else, two dimensions are involved: the functional dimension specifies the system's target (what is it trying to represent) and the semantic dimensions accounts for what T actually represents (i.e. T's representational contents).

For instance, if you draw a map of your city to a friend, your city will be the functinally fixed target, even if the map pictures someplace else due to some innacuracy.

Each dimension (semantic and functional) comprise an independent *explananda* and requires its own theory, for the capacity to have X as a target is distinct from the capacity to forge and exploit a representation of X.

This independence allows us to see that the intentionality of a given state or process is not a product of the exploited representational content, for it comes from the fixed target.

From this vantage point, we can see that much of the current contemporary debate among representational and non-representational frameworks is actually about the strategies through which the system fixes its targets.

With the target/representation distinction in mind, let's see how we can overcome two common antirrepresentaionalist claims: 1) representational contents are inconsistent with naturalism and 2) they have no real explanatory role to play in scientific accounts of cognition.

What are mental representations?

The difficulty in naturalizing representational contents was recently synthesized in Hutto and Myin's "hard problem of content" (HPC): causal correlation does not imply contents, and as that's the only source of natural information, there can be no naturalistic accout of representational contents.

Thus, any semantic gloss given to causal covariation is in the eye of the beholder (i.e. the theorist), and though it may be methodologically useful, it is not doing any explanatory work recognizably representational in nature.

I'm sympathetic to the idea that covariational theories can't ground representational contents, but even if someone eventually manages to overcome HPC, causal covariation would remain unacceptable in RPC because it implies LOT, which we already rejected in the last chapter.

It implies LOT because covariation necessarily relies on a finite set of feature detectors, and so in order to harbor content productivity they need to be aggregated in complex compositions, but compositions of basic elements is the crude definition of LOT.

Fortunately we can bypass any worries with covariation, for it is simply not true that covariation is the only naturalistically plausible source of information: there's also isomorphism.

Cummins' theory of representational contents to account for the semantic dimension relies solely on it: the representation relationship is nothing but the mathematical relationship of isomorphism, i.e. for any x, if x is isomorphic to y, x represents y.

Crucially, the representation relationship does not rely on the fact that it is being currently exploited as such by some system: a map of a city does not become contentfull only when used by someone, i.e. the content is intrinsic and requires neither conventions nor covariation.

Thus, to claim that A represents B means just that A is isomorphic to B, not that A is exploited as a model of B, and only the latter requires B to be fully determined (such determination is the business of target fixation theories).

Cummins' theory dodges HPC for not relying on covariation, so there is no grounding problem.

Morerover, there's no mystery regarding the token's causal efficacy, for both contents and causal powers are products of the token's structure, and the possibility of mismatch between representational content and the system's fixed target enables a robust and purely semantic conception of misrepresentation (to be further developed).

In a nutshell, representations are isomorphisms, and the question of what mental representations are is distinct from the question of how they're exploited within the systems which is an empirical matter.

Do we actually exploit structural correspondences?

Though we can be wary of symbols, it seems clear that human cognitive machinery is capable of exploiting causal correlations through indicator signals (i.e. detectors).

In contrast, the idea that we exploit structural correspondences may not be so readily acceptable, for the mere existence of structural similarities in the brain's activation patterns and wordly structures is not enough to show that it's the similarity that's being exploited.

First example: natural images

A good example of this point can be found in the sparse coding approach to the brain's processing of natural images.

The core idea is that natural images result from collections of indicator signals.

This is not a general principle for image processing, but an empirical thesis, for the indicators come from the statistical structure of the images.

Despite the involvement of indicator signals, images are not simply compositions of them, for the same set of them can account for distinct ones.

Thus, the scheme's minimal semantic elements are not identical to the set of its basic semantic elements, as illustrated by the scheme pallete-c, and this means it cannot be language-like.

The role of indicator signals is not that of providing basic semantic elements, for the actual contents of images involves significant background cognitive knowledge (i.e. complex ways of exploiting the available assets).

The sparse coding model prompted a whole field of study.

The approach was later applied to other modalities such as video sequences.

What's crucial: the image's content is a product of the token's geometry rather than the source of the involved signals, which allows for the kind of structural scheme we've discussed in the previous chapter.

Second example: rats, mazes and food

In order to reinforce the idea that structure exploitation is a feature of biological cognition in at least some cases, we need to show that and how structural representations can also participate in non-visual domains.

Cognitive maps in rats are perhaps the most used example of structural representations in biological systems.

In order to rule non-representational accounts of the rat's capacity, we need to show that representations comprise the best available explanation, specially agains non-representational accounts such as those formulated solely in terms of affordances.

Newen and Vosgerau show that affordaces have more difficulty in accounting for the available evidence (e.g. the rat's learning trajectory and the way information from distinct affordances would have to be articulated).

Third example: causal structures

Let us now consider the structure of causal domains.

There is evidence that in some cases human reasoning relies on exploiting structural correspondence.

In a particular experiment, researchers have shown that we do so in order to find the most rewarding causal path somewhat like a rat finds its path through a maze.

The researchers managed to find some heuristic strategies we apply in dealing with increasingly complex paths, and they amount to a kind of surrogative reasoning that relies on structural models.

How to tell whether one is exploiting representations?

In order to know whether a system actually exploits geometrical structure, Shea suggests the connection between representational accuracy and system performance: if lower accuracy reduces performance, then the behavior is likely to rely on the structural correspondence.

Something like this seems necessary because merely detecting instances of structures within the cognitive machinery can amount to non-exploited contents.

Indeed, for non-innate representational abilities, non-exploited contents are the rule rather than the exception, for every content must be available before the system can learn to exploit it.

Moreover, even detector mechanisms are a potential source of structural information that the system can use to learn about the detected feature (rather than just learning that it was detected somewhere).

So we must capture more than the existence of structure (we have to capture the fact that the system is exploiting it in virtue of its form), but Shea's suggestion is misleading because good performance may come with bad representational accuracy, and bad performance can occur even in the presence of accurate structural resemblance.

The more general notion of incidental effects seems to be what we need.

Distinct models producing the same behavioral output may still have distinctive incidental effects.

The argument in favor of representational accounts of rat's cognitive maps are a real world example: even though the kind of behavior can be explained without representations, this alternative would imply incidental effects that are incompatible with the available evidence (e.g. a distinct learning rate).

A similar reasoning can be used to distinguish cases in which the cognitive apparatus is relying on representations from those in which it relies on encodings, for different schemes provide distinct *systematicity effects* and these amount to distinct incidental effects.

Priming effects are an example of possible systemacity effect: a scheme might result in faster processing of its closest systematic variations.

Indeed, incidental effects are informative even when absent (i.e. when ny two models are able to fully account for the available evidence without any apparent distinction).

The role of representational contents in the assembly of attitude contents

Reliance on structural representations is commited to a somewhat radical departure from classical cognitivism.

To begin with, we must reject Fodor's dictum: to harbor an attitude with the content p is *not* to harbor a representation with that content.

Whenever one says that the system harbored an |x|, this means that the system *applied* a content to a target. It is a case of an *application*, in which both target and content dimensions are fully established.

Applications are not just a new way to talk about classic representations. Unpacking applications is trickier because part of the application content can be due to its target rather than its representational content.

Some applications may constitute a fully-fledged attitude (e.g. a belief), but that need not be the rule.

Attitudes can be much more complex and comprise articulations of multiple applications, attention mechanisms, affections, indicator signals and other kinds of cognitive resources.

The upshot is that representations, targets and attitudes have distinct *explananda*, and whatever our favorite theory of attitude contents is, it need not worry about semantic misrepresentation, for this is already accounted for at the application level.

Non-representationalists also reject the dictum by rejecting any role for representations, but the previous discussion has shown that this is not necessary, for we can instead reconceive representation's role: it need not bear all the explanatory burden of accounting for attitude contents.

In the resulting picture, the contents of attitudes (propositional or not) are not something that scientists will get from philosophers working out foundational issues, but something that must be empirically figured out.

This position is evidently related to a contemporary debates regarding the nature of belief: while Fodor's dictum puts all the explanatory burden of attitude contents in representations, Schwitzgebel's dispositionalism puts none.

That's why it is important to keep in mind that the remainder of this work does not rely on any specific account of attitude contents (except of course for the rejection of Fodor's dicutm).

Thus, whatever representations buy us, it must be specifiable independently of any particular theory of attitude contents. Figuring that out will occupy us in the next section.

What representations buy us

Both representationalism and antirrepresentaionalism are problematic, if we understand them as theses about the essence of cognition.

But we can put this idea aside and argue that structural representations are fully compatible with the core assumptions of 4E cognition: cognition can (rather than needs to) involve the exploitation of representational assets at least sometimes.

The typical source of resistance comes from thinking that, by allowing the entrance of cognitive assets born in classic cognitivism, one is making room for the sub-repticious adoption of unwanted methodological principles, but such principles regard only target fixation, not representational contents.

Representations, alongside computations, detectors, environmental couplings and the like, should always be regarded as a pool of available cognitive assets, and determining which is being employed is an empirical matter.

But there's one additional source of antirrepresentational resistance: the claim that representations cannot bear any relevant explanatory role.

However, that is not so. The independence of the target and content dimensions enables an additional dimension for empirical hypotheses: those relying solely on the difference between what a token represents and the target against which it is applied to.

In other words, we can disstinguish cases in which correct information is being mishandled (using a good map upside-down) from cases in which incorrect information is being properly handled (correcly using a bad map). The latter is an example of a purely semantic notion of misrepresentation.

In order to appreciate this insight, it is crucial to emphasize that representational targets are distinct from actual states of affairs.

Furthermore, as misrepresentation is disjoint from any sort of malfunction or exploitation error, there's no direct connection between semantic innacuracy and unfit behavior.

It is commonplace to give up on accuracy in virtue of tractability (a very detailed map can be more accurate, but it is also more coognitive costly), which means that misrepresentation can be the reason behind apt behavior.

Antirrepresentationalists are probably tempted to disregard misrepresentation as the kind of error that can be reduced to exploitation error, but now they have no HPC to rely on, and they can only bet that non-representationalist accounts will always turn out to be the best available explanations, a much weaker position against representations.

But perhaps surprisingly, antirrepresentationalists are not the only ones tempted to make that move and claim that semantic innacuracies are actually exploitation errors. Most representationalists do that as well, as we'll see now.

What most representationalists got wrong

Classically, in trying to account for the explanatory value of representations, many representationalists stick to what Haugeland called *change of dimension*: a complex causal structure is revealed as semantically assessable.

For any science trying to describe cognitive capacities in terms of the satisfaction of epistemic constraints (partially or completely), this is a crucial move, but dimension change is compatible with accounts that trivialize or deflate the explanatory role of representations in cognition.

Thus, representationalists agree that in order to have representations playing a non-trivial explanatory role, we need misrepresetation, but they usually think that the only possible naturalistic account of misrepresentation involves taking misrepresentation as boiling down to non-representational issues (malfunction, less than ideal conditions, etc.).

To see this, consider a crude version of conceptual role semantics (CRS): the usage pattern fixes both target and representational contents, precluding any possible mismatch between them.

That is why Cummins claims that CRS is a kind of (equally crude) valence theory (in reference to the valence bon theory from chemistry): valences are a kind of fiction.

In other words, within CRS the representation's contribution can't go beyond that of glossing complex causal structures in semantic terms, i.e. dimension shifting, which renders them explanatorily idle (the explanatory burden is being carried by the representation's raw causal role, rather than its contents).

Since CRS conflates the target and content dimensions by definition, in order to avoid trivialization one has to bring some new, non-semantic, element to the table in order to distinguish content-fixing cases from content-determining cases. This allows for mismatch, but it's just like the antirrepresentationalist strategy of reducing semantic error to other kind of error, thus preempting the purchase of a purely semantic dimension for empirical theses.

This conflates semantic accuracy and behavioral fitness, as the norm for correnct representation boils down to the norm for apt behavior.

A more concrete example can be found in Millikan's strategy: the non-semantic element she brings to distinguish accurate from innacurate representation is adaptability.

Thus, semantic accuracy derives from behavioral fitness, but this inverts the explanatory order: rather than explaining cognitive performances with the help of representations, we're attributing representational conttents with the help of behavioral performances. No wonder the antirrepresentaionalist claims representations are not doing any work. They're not.

That's why we can safely claim: it is the absence of a purely semantic account of misrepresentation that opens up the possibility of questioning the explanatory role of representations, and that is what the distinction between the target and representational content dimensions buy us.

We should stick to Cummins's approach, thus. In the resulting picture, whenever representations participate in the best available explanation for some cognitive performance or capacity, there will be no reason to avoid them.

Apes, humans and representational redescription

We've seen reasons to accept that representations can participate in cognitive explanations, but what about representational redescription processes?

In what follows, I'll suggest that representational redescription — as well as encoding, task-embedding and the like — enable the formulation of more refined and (hopefully) fruitful empirical hypotheses regarding the continuities and differences of human and non-human cognition.

The contemporary debate comprises two dimensions: an empirical characterization of the differences between human and non-human cognition, and a characterization of what underlies these differences. First I'll consider the view of Penn, Holyoak and Povinelli (PHP) about the former.

Roughly, the idea is that humans can make judgments relying on information not directly available to perception (we can rely on "unobservable" properties such as weight).

Povinelli's work concentrates on the ape's capacity to handle things like cause, weight or color across perceptually distinct categories.

PHP think that human and non-human cognition can be graphed in a continuum towards a *physical symbol system* (PSS), which is essentially a kind of LOT-powered system: they call this idea the symbol approximation hypothesis (SAH).

They think that we need symbols (i.e. absolute contents) to refer to things that cannot be structurally "pictured": causality, weight, triangularity, leftness, etc.

Thus, there seems to be a tension between RCP (which rejects absolute contents) and SAH, and I think the change that will solve the tension must be made on the SAH side.

What's wrong with the symbol approximation hypothesis

By depicting human-like capacities as PSS-like, SAH leads us to believe that we need two distinct representational systems: a pattern completion one (as the ones we find in neural models) and a symbolic one.

But there' no sufficiently robust reason to think that non-symbolic machinery can't handle what's needed.

Furthermore, there's nothing necessarily symbolic about the human capacities listed by PHP. Describing a capacity as "PSS-like" or "near-PSS" implies assuming what they're trying to establish.

Perhaps an alternative to try to save the symbol talking is by offloading them to the environment, but the difference between the human and other primate's cognition can be neither solely nor mostly out of our heads, and offloading symbols throwns no light on that difference.

By getting rid of the symbol talking, one can realize that what we need are not symbols for |cause| or |weight|, but rather concepts.

Structural representations cannot picture leftness, but they can underlie a lot of knowledge regarding leftness.

The same goes for cause, weight and so on: what's distinctive of the human cognitive capacities involves differences in what comprises the underlying knowledge about these.

Rather than being increasingly symbol-like, the distinction in cognitive capacities may be along the dimensions unfolded in the previous chapters: we may use different schemes, exploit them differently, etc. This enables a more refined talking about discontinuity between ape's and human's cognition.

Some may say we have an alternative to SAH, not a knock-down argument against it, but there are deeper problems.

Even in fully-fledged PSSs, what explains the capacity to handle cause or weight in different perceptual domains is not a symbol, but whatever the system knows about its referent (i.e. a concept). A symbol that refers to cause cannot do the same job as a concept of cause.

A concept of cause in a symbolic system would be akin to a concept of cause in a non-symbolic system, only seriously constrained, for all the knowledge comprising it would be language-like. Why talk about symbols at all?

A possible reason could be that symbolic systems buy us logic-like compositions, but that's misleading, for thinks like negation doesn't apply to concepts (negating the concept of dog doesn't get you a concept of non-dog).

Thus, the problem with the reliance on symbolic machinery — even by means of approximation — as a way to express what's peculiarly human in cognition, is that PSSs are silent about what really matters.

The relational redescription hypothesis

Similar behavior need not imply a similar cognitive apparatus, and RCP is suited to work out the minutia of PHP's relational redescription hypothesis about what underlies the distinction between humans and other primates.

PHP present their understanding of what relational redescription is, as well as his arguments using as example the evolution of social cognition: many think that social cognition must be grounded in the capacity to render theories of mind (TOM).

In stark contrast, PHP think that social cognition can be articulated without TOM.

While the TOM hypothesis explain social behavior by positing the existence of human-like mental states about someone else's mental states, the relational redescription hypothesis claims that social behavior emerged without relying on anything peculiarly human.

It goes beyond social cognition: the crucial claim is that what we regard as "high level cognitive capacities" comprises the human capacity to reinterpret evolutionarily old abilities and mechanisms used to cognize social or causal domains in less-than-general terms.

This is also the spot where the extra degrees of freedom yielded by RCP pays dividends, for the differences can be modeled as the result of reinterpretative abilities that comes from increasingly complex and richly articulated contents comprising concepts.

Domain bootstrapping

Let's sketch a possible account: what does it take for a system to grasp a domain with a higher level of abstraction?

First, remember that structural representations are compatible with both online and offline capacities, and the same representational asset can participate in both.

Any structural representational resource employed in action and perception (i.e. online capacities) is also available to play non-trivial roles in planning, reasoning and thinking in general (and that comes without any worries about FP).

However, not every capacity to negotiate a domain can be regarded as a simulation of lower level perceptual ones, for we do things like playing chess, which comprises a very distinctive kind of structure. How can such a domain come to be?

Andy Clark presents a possible bootstrap story: by associating a portion of knowledge to a public symbol, we can both associate those "objectified wholes" with each other (as we do in "healthy weight"), and as this kind of association goes on, we get new structured domains, and these can be cognized through specific models (maybe using specific representational schemes), that capture the new domain's expressive power and systematicity.

For instance, by spatially separating washed and unwashed itens, we learn to abstract that peculiar property from the item's other perceptual properties.

Proto-versions of this capacity can be found even in apes.

Though PHP question the interpretation of some of these findings, the issues they point out is mostly related to their commitment to SAH, and thus we need not worry (much) about it.

As this capacity is reiterative, much of human's higher level cognitive capacities might come from cumulative cognitive technology.

Many distinct higher order domains may emerge out of processes like this, including those underpinning social practices, intistituions and of course chess.

Thus, the cognitive gap between humans and other primates might be large enough to establish robust differences in cognitive mechanisms, but not to the point where we would have to assume some completely different kind of mechanism (like a proto-symbolic one), and all the evolutionary difficulties that come with it.

Furthermore, Clark's story is consistent with Hurley's picture of "islands of rationality" we used in chapter 1 in order to characterize islands of integrative capacities.

Now let's put it to test: one of the cognitive abilities that PHP describe as exclusively human is that of making similarity judgments by attending to non-perceptual features.

At this point, it might still be puzzling that, at least in the cases where structural contents are all we have to compare cognitive knowledge, what exactly is being compared whenever two instances of the concept of cause from distinct domains are regarded as similar? So let us unpack this a bit more.

Similarity judgments within emergent domains

To see how structures can handle similarity judgments, let's consider Churchland's suggestion that we can understand semantic similarity (or even identity) in terms of neural state space similarity.

Fodor and Lepore challenged Churchland's account of neural semantics claiming that it required unrealistic conditions in order to ascribe the same representational content to any two systems.

Churchland's answer is that representational states are functions of sets of possible positions within the same activation space.

Structural contents can be thus carried by partitions of activation spaces.

Churchland relies on a technique for similarity analysis suggesting that the structures are objective.

Researchers using this analytical tool have shown that even distinct neural models with heterogeneous activation space bear similar geometric patterns of activation.

The upshot is that we need not fully-fledged "logical and abstract" capacities in order to go beyond perceptual stimulus.

Neiher we need fully-fledged linguistic promiscuity: different systems need only to share enough of the structure domain involved in using a given concept and similar developmental trajectories are sufficient to account for that.

What now?

The first important purchase made in this capter is a non-trivial explanatory role for structural representations even in accounts of 4E cognition.

The second is in the considerably large and flexible dimension in which we can place empirical hypotheses about what's peculiar to human cognition.

Now everything we need to make a representational attempt at handling RP without worring about FP is in place.

4 INFERENTIAL PRODUCTIVITY

The major claim of this chapter is that relevance-sensitivity is a cognitive gadget. It amounts to a "cognitive style", a gadget that evolves through cultural evolution and that is culturally inherited. It guides and constrains how the system exploits its own cognitive knowledge. The gadget comprises a finite and relatively small set of templates that are always operative, always "trying to happen". Such templates arise from the representational redescription made by neural control structures (i.e. gating mechanisms) that are deeply involved in the brain's dynamics. Such redescription process involves learning from second-order structures that emerge out of the system's stored experiences (i.e. the set of ways in which the system has previously exploited its own cognitive knowledge). As the dynamics of the system's stored experiences are shaped with the help of culturally established traits, the structure of their dynamics represents what's culturally relevant. While learning from them, these redescription mechanisms apply further culturally inherited biases. This allows the system to forge and exploit representational permutations of its previous experiences in way that is flexible yet constrained by what's culturally relevant. Thus, both the system's experiences and the way they're exploited in order to guide cognition are shaped by culturally inherited traits. This results in a picture where the cognitive machinery's dynamics is both constructed and constrained by culturally established relevancy. We have thus the grounds for relevance-sensitive inferential productivity.

4.1 Where are we at?

It's been a long way, so we'd better take stock and make sure that everything discussed so far is clear. We've seen reasons to think that the capacity to stick to what is relevant is central to human-like cognition and cannot be simply avoided nor put aside. It underlies capacities such as commonsense and situation holism, for without relevance-sensitivity, there's no way to handle non-saturable contexts. The core issue, dubbed *relevance problem* (RP), is to explain how we can know what's relevant at any given situation. What's relevant depends on the system's current context, but figuring out the appropriate context with which to interpret the relationship between the system's surroundings and the system's expectations and goals relies on what it takes to be relevant.

We've also discussed why classic LOT-powered systems seem unable to handle RP: cognition guided by absolute contents face the *frame problem* (FP), an additional issue that renders the pursuit of relevance-sensitivity hopeless. Though one can bet on non-representational approaches to avoid FP, we've seen that the move is not necessary. Structural representations are immune to FP, and this opens the way for an account of RP that does not shun representational explanations away. The deployment of multiple structural schemes enable a non-explosive kind of representational productivity that is compatible with the relational redescription hypothesis. This is a particularly interesting outcome. Throughout the years, RP was regularly characterized either as a problem that would only affect representational frameworks (Bickhard, 2001; Dreyfus, 2007), or as a problem that bothered every framework equally (Samuels, 2010). Neither stance is right on target, or so I've tried to show. RP bothers everyone, but not equally. Those insisting in handling it through absolute representational contents face a dead-end, which is FP. Though the idea that structural representations can be of help against FP is not new (Haugeland, 1987; Janlert, 1996), up to this point, most authors just made a vague appeal to non-sentential representations and left the issue at that point, without fully developing its consequences and possibilities. This of course was never enough neither for a (then) heavily LOT-driven community, nor for those willing to fully reject representations. Costly large-scale overhauls don't happen unless people are convinced that there is no other reasonable way forward. In this sense, Waskan (2006) is an honorable exception: he went to the trouble of trying to show that structures are immune to FP and can account for a reasonable amount of representational productivity. I don't think he went deep enough, though, for I believe that only by going plural about schemes we can see not only that, but also why structural (i.e. relative) contents can really make a difference FP-wise.

Given everything we discussed, it seems clear that figuring out what's relevant cannot be the goal of any functionally individuated cognitive task, whatever the assets involved, for the task itself would inevitably suffer from RP. It follows that relevance must be some kind of emergent property. Functionally isolated performances of distinct cognitive mechanisms and assets bring forth a cognitive economy that mirrors what's contextually relevant. As extensively discussed in the first chapter, the cognitive economy may be understood either in terms of a more-or-less system-wide state or in terms of the articulation of m-contexts. Which story better accounts for what really happens is, and should be, regarded as an empirical issue. That is, the commitment to a cognitive framework or another should not constrain the formulation of empirical hypotheses involving one or another. Thus, whatever the overall system's organization amounts to, what matters is that it constrains the future performances from its comprising mechanisms in a relevance-sensitive way. But of course the million dollar question remains untouched: how can the system either figure out (or be lucky enough) to enjoy the appropriate contextual tonality beyond chance? After all, this seems to assume, rather than explain, some pre-established harmony between what's relevant and the agent's cognitive economy. It remains to be seen how inferential productivity can occur within the contextually relevant set of constraints even though such constraints cannot be cognitively inferred.

The claim that relevance-sensitivity must be a consequence of the system's overall dynamics is specially worrisome for those commited to the idea that cognition can only be properly explained by subsuming performances under some kind of epistemic rationale. The idea behind this constraint is somewhat like contemporary demands of explainability for AI models. Simply pointing out that a model gets there somehow — like most contemporary AI applications do — is not enough. We want to understand its underpinning rationale. Why has a given conclusion been reached? Its inferential process must be rendered intelligible, which means it must answer to epistemic constraints. But what exactly comprises such constraints?

Classic cognitivism understood the pursuit of such a rationale in a very strict sense. It would comprise logic-like inferences expressed in linguistic form. Thus, epistemic constraints were usually understood in terms of truth-conditions. Nowadays, not many are willing to understand "cognitive inference" in such a narrow way. Indeed, if the "ought implies can" principle obtains when assessing cognitive performances of real-world systems, then different architectures, contents and cognitive assets imply distinct epistemic standards under which these must be assessed.[1] If RCP is remotely right about anything, then simulation-based inference relying on structural (i.e. relative) contents is likely to be the rule rather than the exception, at least inasmuch as representational states and processes are involved. Relative contents, as discussed, cannot be assessed in truth-conditional terms. Rather, they're more or less accurate structural depictions. This is definitely not the place to develop such a complex topic, but it is important to understand that the connection between the rationale requirement and cognitive explanation does not imply anything remotely similar to the classic understanding of rationales.[2] Epistemic rationales need not — and probably cannot — be what they used to.

Where that leaves the capacity to handle relevance? It is tempting to suggest that there can be no rationale underlying relevance sensitivity. There are good reasons to believe that no rationale relying on classic computational architectures handling absolute contents can do. Indeed, every rationale seems to suppose it rather than to explain how we get to it. Classic cognitivism would consider anything that cannot be explained under some rationale as non-cognitive. Chapters 2 and 3 have shown that such pessimism is well grounded: given FP, there seems to be no hope for rationales expressed with absolute contents. This is Sperber's approach when discussing sensitivity to s-relevance. In order to avoid circularity, he gives up on rationale-grounded explanations and grounds expectations of s-relevance in non-cognitive physiological markers. But what about relative contents and other architectures, such as that suggested by RCP? A plurality of structural schemes brings with it a plurality of architectural constraints. These are the much discussed assumptions that any scheme makes about its target domain. However, domains are too coarse-grained. They can preempt the system and make sure it won't go astray, that is, it avoids the explosiveness associated with absolute contents. It would be implausible to claim that architectural constraints are enough to avoid RP, though. The major question remains open: couldn't relevance be the outcome of some complex-yet-intelligible rationale relying solely on information structured according to a plurality of schemes?

As it stands, the idea shines no light on how that could be done. We still have to face the challenge of finding the contextually relevant way to articulate the available structural information, which is just another way to state that we still have all the way to go towards a solution to RP. Here's why: in our cognitive lives, we presumably represent and exploit many structural models, and thus the possibility emerges that we can articulate them in indefinitely many ways. This multiplies the set of possible permutations indefinitely. Thus, in order to

[1] See Cherniak (1990) for a development of this point.

[2] See Cummins; Poirier; Roth (2004) for a first step in fleshing out this idea.

handle RP, one has to know which models to apply, and which among the resulting set of possible permutations are relevant. The correct application of structural contents seem to rely on relevance-sensitivity. It is helpful to see how Vervaeke; Ferraro (2013) put this point:

> To direct behaviour towards some future state of affairs requires some representation of that state of affairs. However, representations are aspectual by nature since one does not represent all of the aspects or features of a thing. This means that one needs relevance realization to generate good representations because one has to pick which aspects are relevant to represent. This is a patently vicious explanatory circle. (Vervaeke; Ferraro, 2013, p. 10)

We can understand the "aspectual" nature of representations in the sense that a single object (or process, or state of affairs...) comprises indefinitely many structures. At any given moment, a mechanism can fix on any of these structures as its target. Sometimes distinct temporal and spatial structures can be targeted at the same time and mutually inform each other.[3] But it is evident that the complex articulation of these resources must be already sensitive to relevance. If we try to make it the source of relevance attribution, we fall in the vicious explanatory circle that Vervaeke mentioned. Since this is all evidently true of structural representations, it follows that they provide no direct answer to RP, even though it is an open question whether we could find some rationale leading to it.

Should we go non-cognitive then? The issue with this idea is that whatever accounts for relevance sensitivity must be a potentially exploitable source of further knowledge. We're not only tuned to what's worldly relevant in already familiar circumstances: we can also extrapolate from it. Commonsense, remember, is not just about handling what's routine or typical. Moreover, thanks to the non-saturability of contexts, even typical situations may require atypical connectivity between different portions of cognitive knowledge. Knowledge about time-zones may be necessary to schedule a call, depending on the participants' geographic location. Thus, storing what's typical is not enough. Whatever accounts for out capacity to fix on what's circumstantially relevant must be available for extrapolation in a way that preserves this sensitivity (i.e. in a way that remain epistemically constrained).

This is also manifest in our capacity to make judgments about what is likely to be irrelevant in conservative variations of what's typical. No one is likely to convince a jury of one's innocence by alleging a failure to see the relevance of something widely recognizable as such. Even when providing post-action rationalizations, we can tell the ones that stand a chance in the jury of public opinion from those that do not ("I did not spill the coffee on purpose, I just didn't realize that the cup being full is a relevant feature to consider when deciding to put some more."). Importantly, this is not about being able to extrapolate without failure. We fail more frequently than we'd like to admit. McDermott (1987) exemplifies this with his story of how he was soaking a tablecloth in the sink while his wife was using the washing machine. Both knew the washing machine would discharge into the sink and cause a deluge. Even so, none of them realized the need to do something about it. But when the deluge

[3] This is the core idea behind predictive processing's multi-level architecture.

actually took place, they could immediately understand what happened. The point, thus, is not that we always recognize what's relevant. Rather, there seems to be a properly human way to fail at that while keeping above chance. Some failures may sound funny, because most people can recognize themselves as being able to make the same mistake. At the same time, other failures would be regarded as strong evidence that there's something wrong, or that one's engaging in a childish lie. Thus, accounting for relevance presents us with apparently incompatible requirements: relevance sensitivity cannot be the output of any cognitive task, and yet it must be at least potentially exploitable by other cognitive processes. How can this tuning be sensitive to features that are available from the complex articulations of different mental models without actually making use of them?

In what follows, I'll make a tentative suggestion of how this can be achieved. It relies on structural representational contents and mechanisms of representational redescription, both understood as previously discussed. But it's not my intent to claim that this is the only way to account for relevance. Those working with non-representational approaches (e.g. enactivists) may well provide a similar bootstrap story that does not involve any kind of representational contents. According to the claims previously made, whether a given system relies on representational contents is ultimately an empirical question, and should be determined one capacity — or performance — at a time. However, I do think that structural representations provide the best trade-off between theoretical commitments and cognitive purchases that nature has at its disposal, and that's why I think that a representational account is worth formulating and submitting to further empirical validation.

We'll start with a quick discussion about how cognitive mechanisms can achieve the right level of flexibility that context-sensitivity requires, all without giving up on mechanistic explanations. With a flexible machinery and a place from which a system can learn what's relevant, we still have to face how we can learn what's relevant. Learning is a relevance-sensitive task itself, after all. I'll try to provide a bootstrap story of how we "can learn to learn" what's relevant with the help of Cecilia Heyes' cognitive gadgets framework. If successful, this will be enough to account for how we deal with what's typically and familiarly relevant. But a successful account of relevance sensitivity, one that is able to render commonsense and situation holism non-mysterious, must also explain our capacity to extrapolate from the familiar without loosing track of what's relevant. This will be the last step in the discussion.

4.2 Effective connectivity

How can the brain's overall structure be functionally analyzable, broadly integrated and yet flexible enough to engage in tasks that seem to require a broad redesign of its whole functional structure? This question underpins much of the discussion that took place in the first chapter. On the one hand, we have frameworks that unwillingly preempts any hope. Classic LOT enables a broad integration by adopting the assumption that the mind trades on a single currency, which is the absolute content that LOT carries. But whenever the kind of flexibility required for RP is needed, FP ensues and render the requirement unreachable. The more flexibility is needed, the less we can model it. The attempt to tame it with m-contexts (as discussed in chapter 1, they're roughly a kind of compartmentalization that enables the system to select small portions of its cognitive assets) has leaded us nowhere. A pessimistic view on this is behind Fodor's famous "law of the non-possibility of cognitive science" (1983). In a common variation, proponents of massive modularity usually - but not necessarily - rely on a similar trick. Though they reject the idea of central processing, they must be able to articulate many modules when handling complex tasks, and the positing of a common currency to inter-module communication is surely helpful.

On the other hand, there are frameworks that seem to enable a high level of cognitive flexibility. And they do that without giving up on the idea of m-contexts. The already discussed example is Wheeler's broadly Heideggerian approach (2005). However, there is a catch: whenever the requirement for flexibility becomes too complex or too broad, Wheeler's account relied on a broadly dynamic account of the brain dynamics - one heavily relying on Dynamic Systems Theory. Dynamic models, however, employ a different explanatory strategy, which is subsumption under law (Cummins, 2010a). In contrast, the explanatory strategy being pursued here is the more "classic", mechanistic one: we want to know how something works. We want a broadly functional and mechanistic account of how the mechanisms underlying a given capacity or performance work - the kind of explanation that is revealed by functional analysis.

Maybe Clark (1997a) was right when he claimed that some of the brain's dynamics are so complex that they cannot be subjected to the analytical explanatory strategy. But I'm not so sure. Much of this thinking is motivated by an excessively narrow conception of the analytic strategy. It is usual to think that it implies adherence to classic computational models, symbolic media or even particular architectures such as Hurley's sandwich model (Hurley, 2001).[4] We've dedicated a good deal of chapter 3 to see that none of that is necessarily the case. Rather, it is possible that a structural representation can be exploited by the same kind of coupling that would comprise the exploitation of a structure that is not within the cognitive apparatus. If there are mechanisms capable of coupling with external structures, there can be mechanisms coupled with inner structures as well.[5] And it may well be that such couplings are too complex or too fine-grained to be functionally analyzable. But this goes both ways: if the analytic approach is not committed to the kind of computational architecture nor the representational *genera* that its adversaries take it to be, then it is potentially able to encompass a lot more than it is usually thought. Piccinini's work (2021, 2022) is a good example of how this idea can be fledged. His framework employs the analytic explanatory strategy (in his words, mechanistic explanations) on 4EA cognition. Thus, it is quite unclear whether this explanatory strategy has the limitations that many attribute to them. Most (perhaps all) of the

[4] I take this to be a fundamental issue with the arguments employed by Van Gelder (1995) in his defense of dynamic models.

[5] "There can be" is not an empirical assertion. It means simply that this possibility is not precluded by RCP.

typical examples are only problematic for those commited to the computational architectures and symbolic contents characterizing classic cognitivism.

The main topic of this session remains so far untouched: can a broadly mechanistic approach, at least in principle, make room for the outstanding level of flexibility that seems to be required? A manifestation of the issue, remember, is the question of how a system can have efficient access to the right portions of its knowledge. Functional approaches find the ground for the required flexibility in how the system is functionally organized and compartmentalized. But there's no context-free organization that is adequate to the overwhelming set of non-saturable contexts humans face. Thus, relying on organization implies the need to reorganize and reassemble according to the system's current environment, as well as its needs and expectations. In this sense, the organization itself plays a heuristic role. Cognitive machinery must be broadly integrated while, at the same time, not restricted to a single organization. But how can such wide reorganization takes place if the reorganizing process must itself rely on how the information is currently organized?

The first step is to get rid of a certain traditional way of thinking about the functional explanatory strategy. This view assumes that the decomposition afforded by functional analysis yields a 1-1 mapping between a function and a neural substrate. Each neural structure can carry a single function. Sometimes this is taken to mean that we can make sense of these functions by studying them in isolation. But in many cases, there's no such 1-1 mapping between functional attributions and neural structures Pessoa (2022). In different circumstances, the same neural substrate can carry distinct functions and the same function can be yielded by different neural substrates at different times. The upshot is that there can be more than one way to analise the same cognitive function or performance, for the analysis itself may be context-sensitive.

With this in mind, the second step is to provide a sketch of how could this story about context-sensitive function attribution be told. We can find a clue in what Clark dubbed *neural control structures*. These are:

> (...) any neural circuits, structures, or processes whose primary role is to modulate the activity of other neural circuits, structures, or processes — that is to say, any items or processes whose role is to control the inner economy rather than to track external states of affairs or to directly control bodily activity (Clark, 1997a, p. 136)

The hypothesis is that there are neural populations devoted to modulate the flow of activity between cortical areas. Essen; Anderson; Olshausen (1994) provides a useful analogy: such neural populations are responsible for the inner traffic of useful materials between functionally distinct portions of the brain.

> The production process involves careful selection of useful materials, discarding of excess or unnecessary materials, and transforming and repackaging of the desired materials in an appropriate configuration for the particular applications for which the product is intended. For efficient function, the flow

> of material must be carefully monitored and controlled. This requires specialized systems that are explicitly designed for this purpose, rather than for construction and fabrication processes per se. (Essen; Anderson; Olshausen, 1994, p. 298)

If van Essen is right, there are mechanisms specialized in providing a kind of gating function between distinct mechanisms. By "gating" I mean not just enabling or disabling information traffic. It has an active participation in how different mechanisms in distinct parts of the brain may articulate themselves and influence one another. The exact nature of such mechanisms may vary and there may be dispute about how they're better modeled. Essen; Anderson; Olshausen (1994) take them to be functionally identifiable information-routing control neurons in the brain. Damasio's well-known convergence zones would be another example of brain structure that could fill the role of gating information among distinct mechanisms (Damasio; Damasio, 1994) And of course there's Clark's contemporary approach within predictive processing, which relies on precision weighting (Clark, 2016) All of these comprise wide-scale theories about how the brain is able to integrate and articulate distinct mechanisms in complex ways.

Neural control structures allows us to understand how many distinct mechanisms can interact in ways that render RCP's suggestion of representational productivity more plausible. The involved mechanisms need not trade on a single currency, that is, they can represent or encode information in proprietary formats, but still be integrated and articulated. Friston (1994) calls the influence that a neural mechanism is able to exert over another its *effective connectivity*. We can expand this understanding and say that the set of possible inter-subsystem influences comprises the system's overall effective connectivity. This is distinct from the pattern comprising the brain's physical connectivity, for it may involve lots of short-time transient assemblies comprising very specific articulations of distinct mechanisms. In the resulting picture, distinct mechanisms - perhaps representational ones - with different architectures exploiting distinct cognitive assets may interact in complex ways. This complex articulations of distinct mechanisms enables the system to comprise transient sub-systems that are tailor-made for the situation at hand. Indeed, as previously noted, there is growing evidence that the brain is able to broadly reconfigure its overall pattern of effective connectivity in a fluid and context-sensitive way (Anderson, 2010, 2014; Pessoa, 2022).

This puts a lot of weight on neural control structures — and on whatever specific mechanisms turns out to play this role. They enable us to understand how can broadly mechanistic account can present a high level of flexibility. But how do they "know" the pattern of effective connectivity that the current situation requires? Van Essen claims that, in at least some cases, like that of the visual processing machinery, the involved gating mechanisms are innate (Essen; Anderson; Olshausen, 1994). But of course this falls short of an adequate account. If we are supposed to get off the ground, neural control sub-systems must be developed throughout the agent's lifetime. That is, the agent must learn its way through the possible patterns of effective connectivity. This amounts to the need to learn what's relevant at each situation.

At this point, it is tempting to think that there's no mystery to it. As the agent grows, it becomes familiar with a growing set of specific situations in the world. Connectivity patterns are a just way to "store" this familiarity. In this view, the system's inputs are enough to drive neural control structures towards the right direction. Thus, in similar enough situations, the required pattern will be rendered again and the system will be able to handle it as before. All that neural control structures would need to do is analyze the inputs, search for similar structures and modulate the flow accordingly.

However, similarity-based recognition of the current situation is a non-starter. Everything is similar and dissimilar in open-ended ways. Being comprised of mostly water is equally true of both the human body and a cup of tea. Moreover, both can share lots of typically irrelevant features, such as being in the same place, existing the same time, and having less weight than the Eiffel Tower. Thus, even to recognize something as similar the system must first zero in on the set of relevant features. If we make similarity recognition to be neural control structure's business, the only achievement will be to shift RP someplace else. To make things worse, the same issue applies for learning tasks. The environment in which learning takes place is equally filled with irrelevant features. We have the required flexibility, but it remains blind to what's relevant. How can it learn to see?

4.3 Can we learn what's relevant?

Where does relevance come from? Who or what gets to determine what is relevant and what is not? Fodor's answer to this question draws on scientific inference: relevance is an objective property. It exists independently of any cognitive system. It is no coincidence that Fodor's paradigmatic examples come from science. For instance, most metallic materials expand whenever heated. Thus, changes in temperature are relevant to determine changes in its shape, volume or density. But no one would — at least no one should — understand relevance in such a narrow way. This kind of knowledge is the result of years, sometimes centuries of hard work from multiple agents building on one another. If RP was about tracking this kind of relevance, it would amount to the problem of how we can know everything there is to be scientifically known. That's evidently not what's at stake. Science is likely to be human's most challenging endeavor. The task of figuring out the relevant many-to-many relationships that everything bears to everything else is likely to be a non-ending task. Even though scientific endeavors allows us to know some of what's objectively relevant in this sense, this is definitely not what we have in mind when we claim that human cognition is sensitive to relevance.

Some might be tempted to disagree by saying, e.g. that the relevance we're sensitive to is objective in Fodor's sense, but that it should be regarded as a regulatory ideal. In this view, we're the kind of creature that got lucky to grasp enough of it to ensure the continuity of our species. But that's not the point. Science is about what's true. Relevance, as discussed here, is about what's fit to a given context. And this context is the complex articulation of many environmental and cognitive assets, including the systems' goals and expectations. In this broader sense, what's contextually fit may coincide with what's true, but that's not always the case. It encompasses also pragmatic considerations. Is it OK to answer "would you pass me the salt" negatively? The doorbell is ringing, should I answer it? The norms against which answers to these questions should be measured with are not a regulatory ideal of what is objectively true. Cognition must track what is taken to be appropriate in a given culturally pervaded human world, i.e. the structure of human activities and the set of possible contexts in which they may take place. In this sense, human cognitive activity must be constrained by the structures comprising human activities. It comprises clusters of different kinds of norms and distinct kinds of authorities.

We are introduced to what is relevant as we're introduced to everything that comprises the human world. This may involve lots of things, such as scientific knowledge, institutions, habits, community standards and other cultural elements. As we learn to inhabit the human world, we learn to cope with what's relevant for the activities that take place within. In this sense, relevance emerges as a structural trait that's out there for us to cognize. This is where the temptation to offload relevance-sensitivity to the world takes place: can't we simply let the structures of the world guides us? We need not infer that we're in a certain circumstance. We're always coupled or somehow tuned to the right set of features of the world, and this puts our cognitive machinery in the right contextual tonality (i.e. in the right pattern of effective connectivity). Thus, the human world itself is structured in a way that takes us from context to context. But we must remember an already formulated issue: what comprises the "right set of features"? The same state of affairs can be framed in indefinitely many distinct ways, and these inevitably take into account the goals, interests, and the overall set of available cognitive capacities of the cognizer. To track what's relevant involves much more than simply let oneself be guided solely by what's out there.

Moreover, even if we could offload all this cognitive work to the world, it would not be enough. We quickly addressed this issue in the first chapter: learning is also a cognitive task, and as such, it suffers from RP. As a first approximation of what's at stake, let us consider how the issue arises in artificial neural networks (ANN). What makes them really useful in machine learning is not so much their capacity to recognize particular elements that were present in their training corpus, but the capacity to generalize from them. This kind of generalization is of a different sort, however, for we're not talking about induction. Generalization in machine learning means a kind of tolerance for difference. This capacity underwrites the network's knowledge that the current input is a familiar face under different light conditions, and not a face from somebody else. But it's very hard to learn the right amount of tolerance. If it tolerates too much, the ANN will get a lot of false positives, but if it's too narrow, then the result can hardly be considered knowledge about any structured domain, for what we'll have is akin to a look-up table used to label previously known inputs. This is why the ability to generalize is necessary. But if that's the case, then we have to deal with a pressing question: which features among those that figure in the training *corpus* should be considered when generalizing? This kind of issue is well illustrated in the following charge:

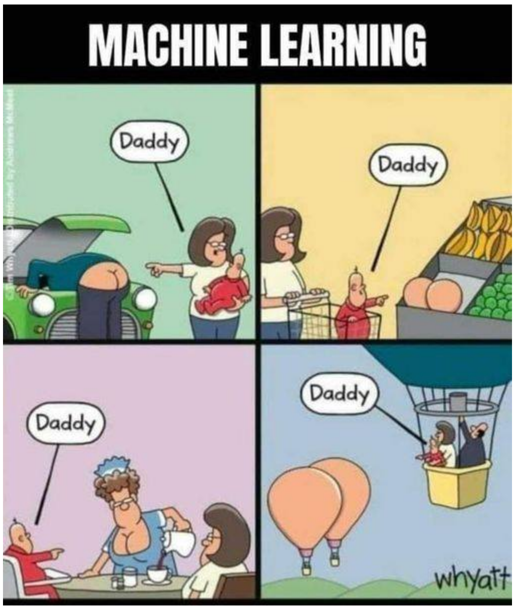

Figure 13 – What's relevant about daddy?

As the picture suggests, a given feature can be seen in lots of different structures, but the feature in question is clearly inessential to what the system was trying to learn about (presumably, daddy's perceptual characteristics). Another interesting example can be found in Geirhos *et al.* (2019): the algorithm used to learn about elephants from a training corpus comprised of elephant pictures, usually takes skin texture as essential. Since skin properties are inessential to elephants, the resulting model knows a lot about elephant skin, but very little about elephants. It can't recognize elephants in cartoons or elephants in lighting conditions that occlude skin details (such as an elephant against the sun). According to the researchers, however, this tendency is easily corrected by tweaking the modeling process in the right way. This can be made both through changes in the training corpus (add, remove or reorganize data) or in the statistical assumptions of the learning algorithm. But if such a tweaking has to be figured out by the theorist, then what are we to say about learning in the wild?[6]

Presumably, there are lots of different domains with many levels of structured relations which need to be figured out when learning how to negotiate them. What can guide learning routines towards the relevant features of whatever it is we're trying to learn about? We can't appeal to the mere availability of the relevant information in the environment, for the question is just how the learning mechanisms can discern relevant from irrelevant features when learning about any given target. Like any well formulated riddle, all the necessary information is *there*, but it is up to us to figure out how things should be assembled in order to get it right. How can we avoid the system from going astray, since virtually any structure

[6] This is not a point against unsupervised learning algorithms. Indeed, the elephant problem emerged in research using supervised algorithms. Furthermore, our learning system is much more robust, for we can learn more or less the same thing out of a wide range of different series of inputs throughout development.

may be (wrongly) taken as essential to the learning target?[7] Notice that there's nothing computationally unfeasible about learning how to recognize elephants from pictures, and still the system might end up taking skin texture to be an essential feature. The problem is not that we lack the computational power to learn all we have to learn, but that in the absence of the proper guidance on how to use whatever cognitive resources we've got, we might still get lost in the way.

In order to dig deeper on the issue, I'll make use of the distinction put forth by Clark; Thornton (1997) between type-1 and type-2 regularities. Intuitively, the idea is simple. Some of our learning relies on being able to grasp patterns or regularities that are only marginally present in the available input information, whatever the domain. As an example, consider the famous mutilated checkerboard (MC) problem. MC is a common example of the role that *insight* plays in problem solving and overall intelligence. The kind of insight that relies on commonsense.

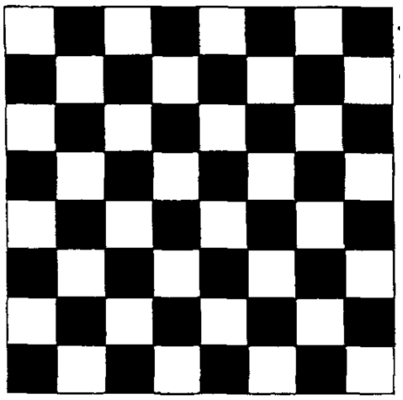

Figure 14 – A normal checkerboard.

Figure 14 shows a normal checkerboard with 8 x 8 (64) squares. Say you have 32 domino pieces. Each one of them can be used to cover 2 adjacent squares in the checkerboard, either vertically or horizontally. The initial problem is: can we arrange those 32 pieces in a way that covers all 64 squares? We can easily see and prove that the answer is yes. However, consider now a mutilated version of the same checkerboard:

As we can see in figure 15, two diagonally opposite corners are missing, which leaves us with 62 rather than 64 squares. Could 31 domino pieces be used to cover the remaining 62 squares? What's interesting about this case is not really the fact that is possible to do it or not (it is not, by the way), but rather, how we can demonstrate it. If we insist on the same strategy applied before mutilating the checkerboard, we'll fall into a combinatorial explosion. Such explosions don't mean a solution is not possible, only that we need a new strategy. And we can easily see that it is impossible by taking into consideration what Kaplan; Simon (1990) dubbed the property of parity: each square has a color, and the squares are organized in a way that takes the color into account. In this case, by alternating them into black/white squares. By

[7] For a deeper discussion of this issue in the context of deep learning, see Arjovsky (2021).

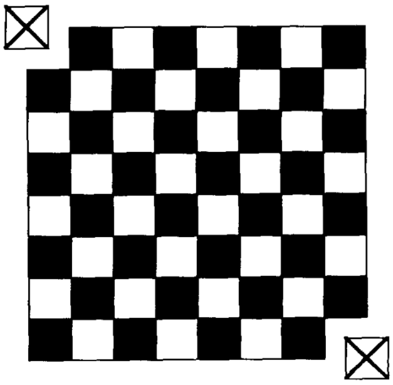

Figure 15 – A mutilated checkerboard.

taking parity into consideration, we can show that every piece of domino will always necessarily cover a black and a white square, no matter how we arrange them over the checkerboard. But figure 15 makes clear that two white squares were removed, which breaks the parity, and allows us to see that a solution is impossible. Given our current purposes, what really matters is this:

> To solve the MC problem insightfully, subjects must switch from an initial representation that considers only numbers of squares and dominos and their geometrical arrangement, to a new representation that takes the parity of the squares into account as well. (Kaplan; Simon, 1990, p. 379)

In other words, the property of parity is irrelevant to solve the first problem, but absolutely relevant to solve the mutilated version. The essential difference is in how the problem is formulated. If we insist in formulating MC as a problem about geometrical arrangement while leaving parity aside, we'll probably not be able to solve it.

At this point, it is tempting to mitigate the significance of this example. After all, the parity is there in the structure for us to exploit. And it seems odd to regard such a clear black-square/white-square pattern to be only marginally present. But this takes things backwards. The point is that the checkerboard has indefinitely many properties other than parity. None of them are equally helpful, but that's something we only know *after* realizing parity's circumstantial relevance. Thus, the issue is to understand how we can fix on parity as a relevant target.

The relation between parity and the geometry of the domino pieces is what Clark and Thornton would regard as a type-2 regularity. Their distinction between type-1 and type-2 regularities relies on the fact that type-2 problems require a rather indirect form of justification. The presence of a type-2 regularity can be inferred from the input data, but not directly. It is there to be found, but only after framing the date it in the circumstantially right way. The following table is used by them to provide a useful illustration of this idea:

Table 3 – Input-output pairs in training set- from Clark; Thornton (1997).

x1	x2		y1
1	2	→	1
2	2	→	0
3	2	→	1
3	1	→	0
2	1	→	1
1	1	→	0

The table presents some type-1 regularities such as `P(y1=1) = 0.5` (i.e. given any input pair, there's a 50% chance of getting 1 as the output) and `P(y1 = 1|x2 = 2) = 0.67` (i.e. given that one of the inputs is 2, the probability of the output being 1 is 67%). These are what Clark and Thornton regard as "direct justifications" that the system can learn about. Such justifications can be acquired by directly inspecting the statistical relationship in the input data.

However, there are statistical relationships of this kind that only become available for direct inspection *after* the application of some arbitrary function. Consider the function `d(a,b) = abs(a-b)`. The output of this function is the absolute difference between `a` and `b` (i.e. negative values are rendered positive and -1 becomes 1). If we apply this function to the input-output pairs in the previous table, this is what we get:

Table 4 – Derived inputs - from Clark; Thornton (1997).

x4		y1
1	→	1
0	→	0
1	→	1
2	→	0
1	→	1
0	→	0

By framing the input set under `d(a,b)`, a whole new set of inferences become immediately available by the same kind of statistical inspection previously available with the raw data. It becomes evident that, whenever `d(a,b) = 1`, the output is also `1`. This means that the data grounds the following conditional reasoning:

```
if x4 == 1:
    y1 = 1
else:
```

```
y1 = 0
```

Thus, assuming that the system is trying to figure out whether the data grounds this reasoning, subsuming the data under `d(a,b)` will make it evident whether that's the case. Clark and Thornton's point is that type-2 patterns become accessible by common statistical inspection only after subsuming the data under some function. I'll call this kind of function (somewhat provocatively) *framing function.* The parallel with the MC problem should be clear already. Unless one subsumes the available data under a framing function that renders parity available, it's very unlikely that a solution can be found, given the combinatorial explosion lurking over every function blind to this property.

On the bright side, this means that many hard cognitive problems from very distinct domains can be solved with a relatively small set of exploitative approaches, many of which (perhaps all) can be shared by both human and non-human animals. Needless to say, many such framing functions can be articulated in complex ways, or even recursively. As Clark and Thornton put it: *"(...) a problem whose mature expression poses an intractable type-2 learning problem can be reduced to a developmental sequence of tractable type-1 mappings"* (1997, p. 61). All one has to do is subsume the available information under the right function. Now, what exactly this subsumption amounts to may vary. There are many candidates among the kind of cognitive tools previously discussed: it may be the use of some appropriate representational scheme, the application of the appropriate exploitative strategy (e.g. task-embedding), or a complex articulation of these. Framing functions can bear both on how we take things to be (e.g. how we represent or encode them) and on how we exploit them. They're a tool for representational redescription.

However, as the authors notice, now we have the problem of selecting the contextually appropriate way to render the available information intelligible among indefinitely many possible ways.

> The number of possible recodings[8] is simply the number of distinct Turing machines we can apply to those data. There are infinitely many of these. Thus the space of indirect justifications is infinitely large. To hit on the right one by brute-force search would indeed be "serendipitous". (Clark; Thornton, 1997, p. 59)

Selecting the contextually appropriate framing function under which to subsume information is essential, and it is also what triggers RP even in regular episodes of learning. Thus, learning processes fare no better than any other cognitive process. This is why one can't simply learn what's relevant. At least not in the same sense that we usually conceive of learning processes. Learning trajectories rely on a fixed target. There has to be something about which one's learning. The mismatch between the system's target and its current guess about the appropriate way to take or handle it, is the only reason why there's some kind of

[8] Clark and Thornton use the word "recoding" to describe the process of subsuming some data set under a framing function. I won't adopt the terminology to avoid confusion with the very specific sense in which the word "encoding" is used here.

error signal to be tracked and minimized.[9] This puts us in a classic chicken-and-egg situation: in order to learn what's relevant about its current circumstance, one has to fix the right set of targets, which means finding and applying the right framing functions in the right order. But doing so takes for granted, rather than explain, the ability to identify the contextually relevant articulation of framing functions.

In learning, a framing function is akin to a developmental tweak. It primes the process towards a certain way of taking the information or towards some structure underlying it. Rather than simply trying out every possible framing function, the system has to already possess at least some minimal information that allows it to narrow down the space of possible functions to try out. But where do these biases come from? And who (or what) gets to say when they are to be applied? Let us consider some traditional possibilities, and the reasons why they're not enough.

4.3.1 Option one: general-purpose learning

The first possibility is that the analysis provided is just wrong. Learning algorithms are perfectly able to figure out how to handle type-2 regularities all by themselves. Or equivalently: the distinction captures no real fact of the matter. There's no need for the selection of the contextually right tweak, for a sufficiently general-purpose learning algorithm is able to encompass every possible context and figure out the right approach, all by itself. Our job is to figure out such "ultimate" algorithm.[10]

This idea is reminiscent of the hope that we can find *the* algorithm comprising general intelligence. This was behind the infamous *General Problem Solver* (GPS) (Newell; Shaw; Simon, 1959; Newell, 1976), which applied means-end analysis in order to build a program that can - as per the project's name - solve anything. The naivety behind the idea was quickly recognized by most AI peers.[11] But the hope never completely faded. We may not be able to build a GPS, the reasoning goes, but we may be able to find a learning algorithm that can build one for us. At least so far, the only general-purpose thing we've been able to find was the "just need more data" excuse for why we're not there yet.

However, if such an approach turned out to be possible, then we would be entitled to suppose that nature draws on the same all-encompassing trick. Thus, it would do no harm to ask: are there any reasons for optimism? Proponents of all-encompassing solutions would be right in arguing that machine learning (ML) algorithms can figure out some problems rely-

[9] Every learning task has a target. Not every learning approach has a supervisor, though. Having a target is different from having a previously established supervisor to help in hitting that target. Thus, one shall not confuse unsupervised learning algorithms, which can fix their own targets at runtime, with "targetless" approaches, which would be nonsense. If there's no target against which to assess the system's current guesses, there's no motivation at all to produce any kind of error signal.

[10] See Domingos (2015) for an attempt.

[11] In 1976, Drew McDertmott wrote: *"By now, 'GPS' is a colorless term denoting a particularly stupid program to solve puzzles. But it originally meant 'General Problem Solver', which caused everybody a lot of needless excitement and distraction. It should have been called LFGNS – 'Local-Feature-Guided Network Searcher"* (McDermott, 1976, p. 4).

ing on type-2 regularities without any further information about the underlying domain. We don't to look for the most recent approaches in order to find this possibility. Even backpropagation learning (Rumelhart *et al.*, 1986), one of the most well-known algorithms even from AI outsiders, can do that. However, this feature may be misleading. The fact that an algorithm can figure out some type-2 patterns doesn't mean there's anything fully generalizable at work. And indeed, backpropagation embodies a lot of assumptions about its problem-space. For instance, if the right framing function cannot be discovered by the gradient descent method it relies on, it will necessarily fail to find it. One can of course bet that the set of all relevant functions is discoverable in this way. But this would not be that different from GPS's assumption that any real-world problem could be expressed as a formal linguistic problem and dealt with by using logic-like inferential tools.

The crucial point is this: what is true of backpropagation seems to be true of every learning algorithm. At least this is what's entailed by the so called "no free lunch" (NFL) theorem, quickly mentioned in the first chapter (Wolpert; Macready, 1997). The NFL theorem shows that every learning algorithm will inevitably embody certain assumptions that will favor some framing functions over others. In the words of Sterkenburg; Grünwald (2021), learning algorithms are model-dependent. They should not be conceived as a one-placed function that maps data directly to conclusions. Rather, they're a two-place function that maps data and some underlying model — which means, roughly, data as framed in a certain way — to conclusions. Thus, the issue with domain-unspecific approaches to learning is not that they can't see anything beyond type-1 patterns. They surely can grasp some type-2 problems, but only inasmuch as the assumptions embedded in the algorithm match those of the domain underlying the target problem. The upshot is that, despite the limited set of cases in which a learning algorithm can successfully figure out type-2 patterns, there's strong reason to believe that their virtues cannot be fully generalized.

4.3.2 Option two: meta-learning

Vervaeke and others suggest that we can handle relevance learning with meta-learners (Vervaeke, 2022; Vervaeke; Ferraro, 2013; Vervaeke; Lillicrap; Richards, 2012). Meta-learning is a promising research field within AI. It targets algorithms that can apply distinct learning strategies and assess its own results, trying to improve whenever possible. Meta-learning systems can be understood in terms of the application of distinct framing functions, as well as complex articulations of them. They have no fixed set of them, nor are they restrict to a single application strategy. Thus, they can try out new approaches and select them whenever improvement obtains, rendering the approach self-adaptive. But the issue is somewhat evident: how can a meta-learner tell which strategy to employ at any given circumstance? How is it to identify situations requiring improvement from those that are fine? It clearly needs to apply its own learning algorithm whose business is to figure out the right framing functions for each situation. But this learning algorithm would embed its own fixed set of biases and limitations as to what comprises the problem space and how it should be approached. Should we perhaps apply a meta-meta-learner? Unfortunately, it wouldn't help, for we'd need a fixed learning algorithm at the meta-meta level as well. The threat of infinite regress is evident.

The authors are aware of this. Vervaeke; Lillicrap; Richards (2012) make a gesture towards a solution by suggesting that meta-learning's business is to become attuned to and handle some of the system's economic properties. The crucial claim is that the meta-learner must rely on opponent processing as a way to balance competing goals that rely on those economic properties. In this sense, it comprises a kind of "virtual governor" that "decides" how to commit cognitive resources.[12] Of course the meta-learner is not to be conceived of as a functionally distinct mechanism, otherwise it would suffer from RP. Rather, the ideia is that what we call "meta-learner" is actually embedded in the overall organization of the brain.

Opponent processing requires the interaction of distinct mechanisms with opposite goals. For instance, there can be a mechanism trying to produce diversification of strategies by e.g. testing new framing functions. Meanwhile, there can be more specialized mechanisms that go deeper by applying domain-dependent strategies the system already knows. At any given context, the amount of resources each of them gets — or how close to a full takeover they're allowed to get — depends on the virtual governor. The *rationale* is analogous to the core idea behind Darwinian evolution: on the one hand, we find the means to produce diversification (mutation, reproduction, etc.). On the other, mechanisms that selectively narrow them down (e.g. competition for environmental resources). The point of course is not that virtual governors have access to every available economic property, otherwise this would quickly lead us to a point where virtual governors and central processing would be indistinguishable. This would bring back the problem of how to select the right approach and the corresponding threat of an infinite regress. Rather, many special-purpose - or at least less than fully general — economic properties result from opponent processing work concomitantly. They're all sensitive to specific elements of the environment, somewhat like the resulting balance of the sympathetic and parasympathetic systems. There's a continuous interaction, and depending on the environmental setup, one gets more time and resources than the other.

In this picture, the capacity to stay tuned to what is relevant is nothing but the appropriate balance that the system obtains from this. Finding what's relevant is understood as the capacity to find the contextually appropriate balance between the many opponent processes going on. Such balance is achieved through the continuous switching between distinct cognitive processes that are always "trying to happen" but get either mitigated and preempted or emphasized and allowed to go on. Thus, framing functions that are too narrow, or too specific may be preempted in virtue of more general ones, bringing diversification to the system. This is somewhat like reaching a point were one would say "nothing is working, maybe I should take this differently?".

This is a nice step in a promising direction. Opponent processing brings to the fore the

[12] The expression "governor" is likely used to resemble the paradigmatic Watt governor from Van Gelder (1995).

limits of conceiving relevance as essentially tied to efficiency.[13] In contrast, the authors emphasize how being sensitive to what's relevant may involve leaving efficiency aside to pursuit diversification. This is what they call resiliency.

> Less explored is the idea that brains are also seeking the opponent goal of resiliency. Brains are trying to maintain an important degree of flexibility so that they have the potential to redesign their function, thereby increasing their fault tolerance in order to retain a potential to resist damage. Thus, the brain can have opponent processing between efficiency and resiliency function as a virtual governor that sets parameters on cost functions that optimize for reward. (Vervaeke; Ferraro, 2013, p. 10)

In a nutshell, the idea is that a meta-learner seeks the best balance between opponent economic properties such as efficiency and resiliency. This looks like a plausible alternative account of the brain's flexibility. By sticking to economical properties, it amounts to a non-semantic and non-computational approach, making it suitable for dynamic modeling. But just like it was the case with Wheeler's CRCs, I don't think they get to the bottom of what's really at stake. The idea allows us to make (some) sense of the complexity undergoing in the brain. It provides some principles we can use to outline its dynamics of reorganization. In this sense, it does throw light on what's underneath the brain's flexibility, but it doesn't really tell us how the cognitive machinery can be tuned to what is relevant. How is the system to know the contextually appropriate balance? What should and what shouldn't get the system out of it? Meta-learners are supposed to figure that out somehow, but if left on their own, they have no clue on how they can do that.

What we have here, I think, is another attempt to stop a vicious regress of explanations for RP by appeal to a change in the explanatory strategy. This is just what we've discussed in the first chapter, when assessing Wheeler's CRCs. Opponent processing may be a nice way to describe the system's dynamics, but it won't help us to understand how does the system work. Moreover, while discussing patterns of effective connectivity, we've seen that there's no reason to give up on the analytic explanatory strategy. At least in principle, nothing preempts the possibility of a sufficiently flexible system that is broadly functionally analyzable. Vervaeke seems to be aware of that, for he recently co-authored a paper arguing that the opponent processing approach and the PP framework are actually two ways of describing the same underlying process (Andersen; Miller; Vervaeke, 2022). What he describes as the attempt to balance competing goals such as that of being efficient and resilient is equivalent to what PP accounts for in terms of modulation by precision-weighting. RCP is not a version of PP processing, but it surely employs the same explanatory strategy and it is broadly compatible with PP mechanisms employed in accounting for flexible patterns of effective connectivity — i.e. the possibility of multiple shades of contextual tonality. At least so far, no light has been thrown in how the system learns its way through what's relevant.

[13] The framework in which this occurs more explicitly is Sperber's. As discussed in the first chapter, Sperber's approach relies on the idea of s-relevance (Sperber; Wilson, 1995), which is essentially a way to measure the level of efficiency that an input enables the system to reach.

4.3.3 Option three: innate tweaks

A different approach involves heavy reliance on innate framing functions. This is advocated by researchers like Gary Marcus (1998; 2003). In this view, we are biologically inclined to handle inputs in a characteristically human way. We're born with the appropriate set of tweaks, i.e. the relevant developmental tendencies (biases, if you will) are genetically inherited. As previously presented, Marcus' view is commited to more than that, for he thinks that some of this innate tweaks develop into symbolic machinery, but this need not worry us right now.

Innate developmental tendencies and tweaks are undoubtedly plausible in some cases. The reason we don't stick to elephant skin when learning about elephants might well be the availability of some innate resource. Likewise, innate developmental tendencies might be enough to explain why a sudden acute sound catches more of our attention than a continuous background noise. But when it comes to social creatures like us, this falls short of what we really need, for it leaves a huge set of culturally established learning targets unexplained. We most certainly do not inherit developmental tendencies to learn the right way to behave in classrooms or in airplanes. We could try to build those targets from genetically inherited primitive targets, but even though the system has the capacity to stabilize on some specific target, it is hard to understand how could such genetically inherited targets be assembled in the right way, so that the systems locks on the relevant learning target dictated by the present circumstances. Furthermore, we'd be stuck with the need for a huge amount of specialized learning mechanisms and/or very specific learning biases, and this would lead us back to square one, for now we'd have to deal with the problem of how to select which learning routine is the right one in each situation.

4.3.4 Option four: incremental learning

While presenting the distinction between type-1 and type-2 regularities, Clark and Thornton advance their own idea about how we learn to apply the contextually appropriate set of framing functions. The core of their suggestion is incremental learning. Hard type-2 problems must be factored in developmental sequences of easier type-1 problems. Rather than trying to figure out what an elephant is out of the raw pattern, the system can initially focus on smaller and easier problems, such as figuring out textures, shapes, relative sizes, and so on. Once it has such chunks at hand, they comprise a set of "primitives", and the problem of whether and how these can be articulated into an elephant becomes achievable through simple type-1 processing. In a nutshell, *"(...) what looks like type-2 learning is in fact the occasional reformulation of a type-2 problem in terms that reduce it to type-1."* (1997, p. 64).

Unfortunately, as it stands this is just a reformulation of the problem. Those more familiar with contemporary neural models within cognitive neurosciences and AI (specially deep learning) will recognize this as a standard way to interpret what virtually any deep neural model does. For instance, models built for handwritten digit recognition (such as those trained against an MNIST *corpus*) are usually taken to employ this approach: first, they recognize small patterns and later they "build" the digits from them. How is this possible? Contemporary deep learning models can do that because they're domain-bounded. Such boundaries are indirectly established by their training *corpus*. They may either comprise a narrow topic (visual features of elephants, digits, etc.), or they may involve a lot of information from distinct topics, but under a single encoding. In the former case, being domain-specific means a larger probability of finding the right set of framing functions (though there's no guarantee, as the example of elephant skin reminds us). In the latter, the system may handle information from multiple domains, but it will frame all of them indistinctly, e.g. whether you're asking it for the name of a country or the result of a word problem, it will handle both prompts by making next-word prediction. Contemporary large language models (LLMs) such as GPT-3 or GPT-4 rely on this (McCoy *et al.*, 2023). They encode the structure of the probabilistic distribution of formal linguistic elements, whatever they're about, and any other structure remains out of reach.

The corollary is that contemporary ML techniques won't buy us what we need. Either we have a set of framing functions bounded to a specific domain, or we have a set of framing functions that tries to handle every domain equally, without the proper treatment of its particularities (In RCP, the latter amounts to the claim they encode — rather than represent — everything under a single structural scheme). Clark and Thornton are aware that this kind of limitation applies to their view. Nevertheless, they partake of a cautious optimism regarding the possibility of reusing framing functions. Indeed, in the second chapter, we've discussed many possible ways in which knowledge about a domain can be employed to exploit another one: relations between schemes, encodings, task-embedding and so on. But we've also seen why that's not enough to avoid RP. If we introduce the possibility of increasingly complex articulations of knowledge from distinct domains, the threat of a combinatorial explosion returns, and with it the need to know in advance the relevant articulations from the idle ones.

I hope that, at this point, it is easy to see how close we are (again) to the kind of issue we've faced when discussing frame systems in cognitivism and SPACs in Heideggerian cognitive science. Learning itself seems to rely on the application of cognitive features in a relevance-sensitive way, which means we're trapped again: we can't simply learn what's relevant because relevance is necessary for learning. Learning helps us realize what is typical, identical, similar and so on. But all of these require context-sensitivity. Even identity and similarity judgments requires one to pay attention to the right set of features. The ones taking the texture of elephant skin as a necessary member of this set in all possible contexts will fail to recognize elephants in many typical circumstances.

Have we exhausted all the options? I'm inclined to think that, up to a few decades ago, we would be forced to say so and leave the matter open. But there's a road that, despite not being exactly new, is getting increased attention and, as far as I can tell, has never been properly exploited as a potential way out of RP. A first version of it can already be found in a speculation made by Clark and Thornton:

> It may even be useful (though clearly highly speculative) to consider public language and culture as large-scale implementations of the same kind of strat-

> egy. Language and culture, we suspect, provide exactly the kind of augmentation to individual cognition that would enable uninformed learning devices to trade achieved representation against computation on a truly cosmic scale. Public language may be seen as a ploy that enables us to preserve the fruits of one generation's or one individual's explorations at the type-1/type-2 periphery and thus quickly to bring others to the same point in representational space. Otherwise put, we can now have learning trajectories which crisscross individuals and outstrip human lifetimes. (Clark; Thornton, 1997, p. 64)

Despite the speculative gloss, relatively recent empirical findings enable us to flesh out a significantly less speculative story. One in which cultural elements that involves language — but is not being limited to — may play a prominent role in explaining how we can handle the relevant framing functions without raising RP. These findings can be synthesized in Cecilia Heyes's framework of *cognitive gadgets.*

4.4 The cognitive gadgets framework

As things stand, we already know that in a very general and explanatorily trivial sense, we learn to recognize what's relevant as we learn to cope with the world. However, as we try to make sense of this capacity at the sub-personal level, things remain quite challenging. The difficulties emerge when trying to account for both how information about the world is stored and how the information is exploited. This is not surprising, for we store what we learn, and to learn is to exploit information yet to be stored. That's why appeals to how the information is organized are hopeless. All they do is to shift the explanatory burden to learning processes. And as discussed, learning seems to be in a seesaw between general-purpose mechanisms that are blind to domain-specific regularities and domain-specific learning that are useless outside their target domain. Clark and Thornton know this well:

> The trouble with informed search, of course, is identifying the informant. In many cases, positing better-informed search simply begs the question. Just where did those feature detectors, or those biases towards trying such and such a recoding first, come from? Unless we are comfortable with a very heavy-duty nativism and an amazing diversity of task-specific, on-board learning devices, we will hope in addition to uncover at least a few more general strategies or ploys. Such strategies (tricks, ploys, heuristics) cannot be problem specific, since this would be to fall back on the full nativist solution. Instead, they will constitute general techniques aimed at maximizing the ways in which achieved representations can be traded against expensive search. They will thus maximize the chances of a learner successfully penetrating some random type-2 domain. (Clark; Thornton, 1997, p. 63)

What they regard as "informed search" is just exploiting information with the right set of tools, that is, information subsumed under the contextually appropriate set of framing functions. But perhaps there is a way to dismount from the seesaw.

"Cognitive gadget" is how Cecilia Heyes calls cognitive mechanisms that are culturally, rather than genetically, inherited. The name is suppose to draw a contrast with what Steven Pinker once called cognitive instincts (Pinker, 1994). For Heyes, much of what we typically regard as domain-specific innate cognitive mechanisms (i.e. "cognitive instincts") are actually gadgets: theory of mind, imitation, causal understanding, meta-cognition, empathy and of course, language (Heyes, 2018a, 2018b; Heyes *et al.*, 2020). Whether she's right about each one of these is of course an empirical matter that's outside this work's scope. But she does present considerable evidence for all of them. Some of these capacities are specially interesting for our current interests: the idea that theory of mind is a cognitive gadget is broadly compatible with the relational redescription hypothesis outlined in the previous chapter under RCP's lights. We have thus an example of the overall compatibility of RCP and the cognitive gadgets framework right from the start. Furthermore, it renders plausible the idea that framing functions can be culturally inherited. As an example, causal understanding — in Povinelli's jargon, the capacity to render `f(cause)` — can itself be a cognitive gadget that is culturally inherited in the form of a developmental trajectory of culturally established framing functions.[14] The idea has the potential to throw light on what makes us human. This is a nice first step towards what will be the major claim of this chapter: relevance sensitivity is a cognitive gadget. But before getting there, we have to outline at least the core tenets behind Heyes' framework.

Cultural evolution is usually though of as a way to account for culturally sedimented information regarding adequate behavior, practices, traditions and religious belief systems. In this picture, such cultural information becomes available to be handled by us with our set of genetically inherited tools (i.e. our cognitive instincts). Culture defines what we'll handle, and genetic defines how we'll handle it. Heyes goes beyond this traditional picture by claiming that we can culturally inherit also the set of techniques we employ when handling cultural information. We don't inherit just the *what*, but the *how* as well. This is what she has in mind when she claims:

> (...) cultural evolution has the potential to explain the adaptedness of distinctively human cognitive mechanisms; why, in some cases, cognitive mechanisms seem to fit the environments in which they operate and do their jobs reasonably well. (Heyes, 2018a, p. 37)

Heyes provides a selectionist account of cultural evolution. The mechanisms responsible for producing variation, as well as those responsible for selecting the most advantageous traits are autonomous in relation to genetic evolution. This means that cognitive gadgets may change solely due to cumulative social learning. Now, if any such account is to get off the ground, it must start by answering three core questions: 1) what exactly is evolving? 2) By means of what kind of mechanisms? And 3) what are the routes of inheritance? Heyes is particularly clear about all of them. 1) What evolves are mental mechanisms functionally individuated; 2) The mechanism of replication is that of social learning; and 3) the routes through which these may happen can be vertical (as when a kid learns something from her biological parents), oblique (as when someone learns from her uncle), horizontal (siblings, friends and so son) and of course any combination of these.

[14] This suggestion can be found in Heyes (2018a).

Social learning is sometimes modeled as analogous to genetic replication. For instance, one can learn to avoid dangerous food by noticing that someone else is avoiding it. In this view, a small information fragment is being transmitted to someplace else. But very few episodes of social learning have this form. Children don't learn how to read by having the right information packets directly copied within their heads through visual or auditory channels. Instead, they're continuously instructed and encouraged through time. This is how Heyes understands the possibility of inheriting cognitive gadgets:

> This insight opens up the possibility that features of cognitive mechanisms (mills) are culturally inherited in the same ways as ideas and behaviors (grist). A cognitive mechanism certainly is not a pellet of information that can be copied inside your head, sent through the air, and planted wholesale in my head. But if grist is not "copied" in this sense, there is no reason to expect mills to be "copied," either. Instead, we can recognize that certain kinds of social interaction, sometimes with many agents over a protracted period of time, gradually shape a child's cognitive mechanisms so that they resemble those of the people around them. (Heyes, 2018a, p. 44)

The idea that cognitive gadgets are a product of cultural evolution inherited through social learning enables a reasonably detailed account of what's distinctive about human cognition. We need not postulate a huge gap in the nature of the underlying genetically inherited mechanisms. Of course there must be something peculiar, otherwise we would be left with the problem of explaining why apes and similar creatures don't share our capacities. But rather than concentrating on peculiarly human "cognitive instincts", we can instead focus on accounting for what's peculiar about human culture. Heyes' account has the potential to supply important additional pieces to the ideas worked out in the previous chapter. When discussing the relational redescription hypothesis, it was stated that structural contents and representational redescription yields a rich dimension of empirical hypotheses regarding what's peculiarly human. But if Heyes is right, this dimension need not be solely populated by genetic mechanisms. Rather, it may involve many products of human culture. This means that, if we ground what is peculiar to human culture in genetics, we will already be preparing the ground for a way of explaining the presence of characteristically human cognitive mechanisms. There might be a peculiarly human way to realize representational redescription, or a peculiarly human amount of redescription our cognitive apparatus can engage in. Both characteristics can be explained by how human culture impinges on it.

Heyes' proposal is an example of how such a story could go. She claims that humans' genetic starter kit have three peculiar traits: first, a considerably higher sociability. Our temperament allows us to tolerate, seek and enjoy one another's presence much more frequently and for longer periods. In contrast, if you put apes within an airplane, a blood bath is a bit more than likely (Hrdy, 2011). Second, we have enhanced social motivation. Innate attentional biases dislodge our attention towards other agents. There are, for instance, biases for peculiarly human voices and typically biological movements. This enhances drastically the power that other minds have over ours throughout our cognitive developments. Third, our general-purpose information-processing machinery is considerably more powerful. We can process more in less time, and store more information as well.

The list is not just surprisingly small, but the nature of the traits are curious as well. The only distinctive developmental biases are those that enhance our sociability. In Heyes' view, we need not posit any innate special machinery, symbolic or not. All we need is an increased level of processing power and hungriness for culturally established stuff. This looks somewhat radical, but it need not be regarded as an all-or-nothing issue. The point is not that distinctively human cognitive tricks must be a product of cultural evolution. What Heyes' framework buy us is a plausible alternative to purely genetic accounts.

In order to put such a heavy burden on social learning, Heyes introduces a peculiarly human kind of social learning she calls cultural learning.[15] Cultural learning is social learning that involves cognitive processes specialized for cultural inheritance. Heyes provides no clear definition of what cultural learning is, but she provides a well-developed list of paradigmatic examples: selective social learning, imitation, mindreading and of course language. In her bootstrap story, these are cognitive gadgets that we develop and acquire through our distinctive genetic starter pack. The gadgets themselves enable further extensions thanks to the cultural accumulation they help to afford. Indeed, the possibility of cumulative culture is a human trace (Tennie; Call; Tomasello, 2009). These cultural gadgets are said to enable cultural learning when they enhance the fidelity with which social learning happens. Such faithful transmission over generations enables a high level of stability, and human culture can thus accumulate modifications over time without necessarily relying on genetics.

In this picture, the claim that cognitive gadgets are products of cultural evolution amounts to the claim that they were shaped by cultural group selection. As an example, consider a case in which people from two distinct social groups A and B are genetically alike but have different versions of the same cognitive mechanism. The version from people of the group A may be more effective in fulfilling some function and achieving some goal (getting food, providing protection, etc.). Thus, inasmuch as that mechanism is involved, this means that people from group A will achieve greater success. This translates into a higher number of descendants and a higher survival rate, which means that group A tends to prevail over B over time. Moreover, a higher success rate my attract people from group B into learning from group A. All of this contributes for the slow fading out of group B's version of the gadget, which means that group A is the fittest.

A nice - and somewhat surprising - example of how a cognitive gadget can develop out of the presented starter pack is the human capacity to imitate. Imitation is traditionally regarded as a cognitive instinct that plays an important role in social learning. Underneath the capacity to imitate, however, we find the *correspondence problem.* Roughly, what looks like

[15] The name "cultural learning" is not really new in the literature. However, Heyes advances a specific way to understand what cultural learning is, and uses this understanding to discuss the role of cultural learning in what makes us distinctively human. Here I'll present just what's indispensable for our considerably narrower goal. For more details, see Heyes (2018a).

imitation from our perspective seems completely different from the perspective of the agent doing the imitation. Heyes exemplifies this with the following illustration.

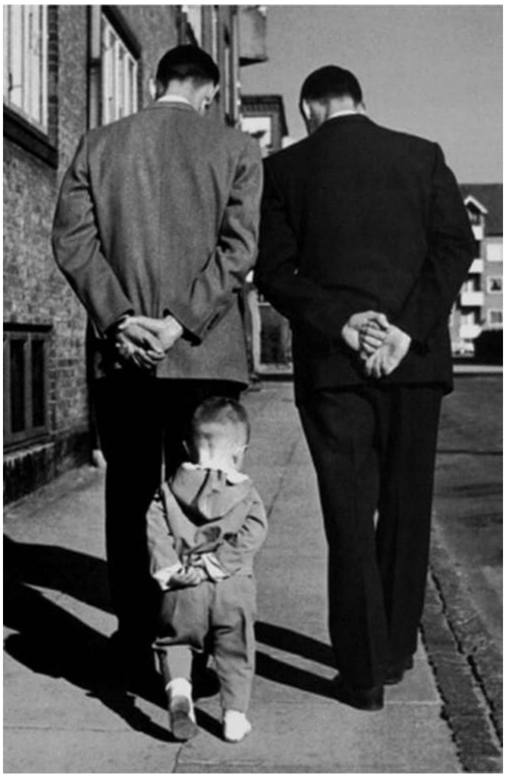

Figure 16 – Imitation in children.[16]

What we see in figure 16 is a child with her arms, head and pace similar to the two walking adults. However, from the little boy's perspective, we have something very different:

> The little boy cannot see the resemblance. As he imitates the men, dipping his head and putting his hands behind his back, all the boy can see is the ground and perhaps the heels of the men in front. To the boy, hands behind-back looks like groundandheels. The boy can feel his own movement— a slight tension in his shoulder joints, the touch of one hand on the other — but he cannot feel the movements of the men. And he can't compare his own movement with that of the men by listening because hands behindback does not produce a distinctive sound that is detectable to the human ear. So, how is the boy able to produce an action that resembles the men's action from a thirdparty perspective when he, the boy, cannot sense — see, feel, or hear— the resemblance? (Heyes, 2018a, p. 118)

The problem, in a nutshell, is how the cognitive apparatus can associate the movement the agent produces with the movement from third-parties that the agent only sees. Notice that appealing to innate and genetically inherited mechanisms provides no complete answer, for we would still own an account of how that mechanism works. That's why appeals to, for instance, mirror neurons won't do. Despite the evidence of their involvement in imitation, involvement is not enough, and now we're left with the problem of explaining how can neurons can provide

[16] The picture was used by Heyes (2018a), but was originally downloaded from https://www.catholicgentleman.net/2015/01/im-just-like-daddy/.

a solution to the corresponde problem (Catmur; Walsh; Heyes, 2009). The correspondence problems shares a trait with RP: in both cases, it's very easy to suggest a solution that just begs the question. Heyes suggests that the solution is culturally, rather than genetically, inherited.

Her idea is that we can associate sensory and motor states bidirectionally. Whenever we produce a visual image of somebody making a movement, we also tend to activate a neural state involved in producing that same movement. These associations are not innate, though, but rather a product of sensorimotor learning. This is what she calls the associative sequence learning (ASL) model. According to ASL, we learn an increasingly larger repertoire of associations between what we see and what we do. This happens in many ways. For instance: babies can learn to associate the visual stimulus of hand movements and the corresponding motor processes by self-observation. This much is also available for non-human animals, but humans have very peculiar additional sources of learning targets. We're frequently exposed to optical mirrors where we can observe ourselves, drastically increasing the opportunities for self-observation. Moreover, babies are frequently imitated by others. There's evidence that western mothers imitate their infants roughly once every minute (Uzgiris *et al.*, 1989). For instance, when the mother imitates the baby frowning, the baby can learn to associate her actual motor state (frowning) with that visual stimulus from the mother. Finally, due to our increased level of sociability, we engage in lots of synchronous activities like rituals and dances.

Thus, the typically human environment provides plenty of opportunities to acquire visual-motor associations of the necessary kind to ground the peculiarly refined way in which humans are able to imitate. Furthermore, this enables us to encode relatively simple body movements into "chunks" that can be further articulated in complex ways. Such chunks can underlie our distinctive capacity for activity coordination, including those underlying communication through gestures. They are a nice example of what framing functions can look like in the real world. Indeed, there's evidence of cultural differences in how facial expressions are processed. Jack *et al.* (2009) found that Western Caucasians distribute their attention more or less equally among the eyes and the mouth, while East Asians tend to concentrate on the eyes. Among other possible effects, this renders distinct ways to recognize emotions in other people's faces. Heyes' development of the possible outcomes is worth the quoting:

> This kind of cultural tuning of attention to social stimuli could have profound effects on social learning. For example, observational conditioning is a primary means of learning the value or emotional valence of types of object or event; people, animals, plants or practices become attractive or aversive when they are paired with positive or negative expressions of emotion by others. Therefore, if Western Caucasians are more sensitive to expressions of fear and disgust, it is probable that they would learn more readily via observational conditioning that certain objects are threatening or repulsive. In this particular domain they may be faster social learners, not because they have better or different genetic adaptations for social learning, but as a result of sociocultural experience tuning input mechanisms to a particular configuration of facial features. (Heyes, 2012, p. 2185)

Though the example is not focused on imitation, it helps us to see how the human capacity to imitate may employ the same kind of sensorimotor learning that is available for non-human animals. The crucial difference is that we apply those mechanisms to the outputs of culturally embedded developmental trajectories in a much larger scale. Notice that this is pretty much Clark's story of how the ability to handle type-2 problems can be decomposed into a developmental trajectory of framing functions that render the information tractable by type-1 learning mechanisms. Only now, thanks to Heyes' framework, we have a plausible source for culturally shaped framing functions and developmental trajectories.

We can make this point even more clear by considering the following question: there is evidence that non-human animals learn by imitation. But if imitation is a cognitive gadget, how can we explain the fact that humans are distinctively good at it? That is, if non-human animals' set of innate cognitive tools is just like ours, how can this apparently huge gap be explained? Heyes' ASL models puts the burden on socio-cultural experiences such as the opportunities for synchronous actions, interactions with optical mirrors and, in the case of babies, being imitated. This kind of social experience is not available to other primates. An important exception might be self-observation, though. Non-human animals can employ the same strategy and learn to associate perceptual stimulus with motor states or processes. But their environment is not so rich as the one available for humans. We have thus a plausible bootstrap story in which a higher level of sociability and more powerful general-purpose processing and learning mechanisms can render a number of cognitive gadgets inaccessible to non-human animals. Imitation is one such gadget.

Well, what else is in Heyes' story for us? Remember that we want to dismount from the seesaw of specialized and general-purpose learning mechanisms. General-purpose uninformed learning is a dead-end, for it inevitably suffer from RP whenever it tries to pick the right framing function. In their turn, specialized learning mechanisms can only work within their target domain. Furthermore, up to this point, the only plausible source for the tailor-made character of such mechanisms was genetic evolution. But these are not fit to explain developmental trajectories and biases that only make sense within the current setting in human culture. As Heyes exemplifies:

> If, as behavioral ecologists and economists have assumed, social learning strategies were fixed products of genetic evolution, they would make social learning selective, but only in a way that was efficient in ancestral environments. For example, if older individuals tended to provide more reliable information in the distant past, agents alive today would be inclined to copy older individuals even if, in a tech savvy world, younger individuals tend to know more, at least on certain topics. (Heyes, 2018a, p. 110)

In the resulting picture, culture is not just a pool of resources (abilities, knowledge, artifacts, etc.) that genetically established mechanisms can learn about. It is also the source of learning tweaks and targets that enable us to exploit all of these resources as well. Culturally inherited learning biases are culturally inherited framing functions. They handle our economy of attention and provide targets against which our learning trajectory can be measured. By providing the right set of framing functions at the right developmental stage, culture is able to provide the right developmental sequence of tractable type-1 learning problems. Type-1 problems, remember, are the kind of issue that general-purpose learners can handle well. Therefore, Heyes' framework buys us a non question-begging bootstrap story that greatly enhances the viability of general-purpose learning: there's no need for them to figure out everything by themselves. They need not select the right set of framing functions by solely relying on a somewhat miraculous genetically sedimented approach. The same learning mechanism can lead to different outcomes in distinct cultures. We have thus a plausible story that enables us to see how our genetic starter-kit can lead us through very similar developmental trajectories without the need of native special-purpose learning systems.

4.5 Relevance sensitivity as a cognitive gadget

Our current goal can be summarized like this: we want to show how a system can track what is relevant and use it to constrain its own activity, even though this is not the goal of any of its mechanisms in particular. This ability to track and constrain must be kept intact as the system switches from context to context. Otherwise, it won't help us to explain relevance-sensitivity in non-saturable contexts. Heyes' framework of cognitive gadgets can help us with that in three steps: first, we have to understand how what's culturally relevant gets stored in a way that enables patterns of effective connectivity (*ec-connectivity* from now on), resembling what's culturally relevant. Second, we must make sense of how what's stored may guide or constrain cognitive processing in the appropriate way. Third, we must know how the system can extrapolate from the contexts it is familiar with without losing track of what's relevant. Let's concentrate first on the question of how what's relevant can be stored within the system. What's stored is what's learned, but as we've seen when discussing the possibility of learning what's relevant, providing a direct answer to how we learn is harder than it looks. Heyes' cognitive gadgets provide a promising path towards a way out, but there are some pitfalls that we must avoid.

As we dig into the human world, we experience and learn about an increasing number of situations and domains. But we don't experience nor learn about distinct situations in isolation. It would be misleading to claim that we experience a situation s1 *and* a situation s2 if they temporally succeeded one another. Rather, we should say that we experience s2 *as coming from* s1. In other words, our experience comprises the dynamics of going from s1 towards s2. Now, given that every situation is experienced through the lens of our cognitive apparatus, we should say instead that what we experience is going from a contextual tonality c1 to a contextual tonality c2. But both context and situation are concepts that involves wordly states of affairs. In what follows, however, I want to concentrate on the overall state within the cognitive apparatus. In other words, I want to focus on the system's ec-pattern. Each of the aforementioned contexts involves such a pattern, so we can say that the contextual tonality c1 involved the pattern p1, and so on.

Now, I want to remark that the dynamics comprised of some sequence of ec-patterns may constitute routines. For instance, every day we may engage in a dynamics comprising the patterns p3, p4, p3 again, p45, p4 again, p3 again, p9 and p20.[17] As the dynamics is reenacted day after day, it may reinforce the connection between the comprising patterns within the cognitive apparatus, to the point where it makes sense to talk of it as a template for the dynamics (call it t1). Thus, t1 = { p3, p4, p3, p45, p3, p9, p3, p20 }. It is a "template" in the sense that one should not regard the dynamics of ec-patterns to be an exact reconstruction of each contextual tonality. Say t1 describes a routine of brushing one's teeth, having breakfast, brushing one's teeth again, having lunch, teeth-brushing again, having dinner and going to sleep. Sometimes the agent may wake up in a good mood, or tired, or perhaps feeling cold. Brushing one's teeth in a good mood involves a contextual tonality that's different from brushing one's teeth while feeling tired. But as far as t1 is involved, all that matters is that she's brushing her teeth as usual. Thus, inasmuch as p3 comprises an ec-pattern, it points towards the set of cognitive assets typically involved in brushing her teeth. The connection among the assets need not be adaptive nor useful. One can conceive, for instance, that every time the agent engages in p3, some fond memories become more salient or prone to come to the fore.

Three quick yet important remarks must be done. First, templates of dynamics in this sense need not arise only from routines. They may be adopted through social learning as well: "You do t1? You should do t32!".

Second, a dynamics like t1 is not supposed to encompass everything that happens throughout the agents' day. Most of us don't live like Sisyphus, i.e. under a single dynamics of pushing a heavy boulder up a hill, seeing it roll backwards, chasing it and starting all over again. A lot happens between brushing one's teeth and having lunch, for instance. What a dynamics like t1 is supposed to express is akin to a cyclic tendency. As the time for lunch approaches, the agent tends to approximate with increased accuracy the ec-pattern involved in having lunch. But other dynamics may come together at the same time. Later we'll see how the suggested approach enables both ways of context switching: one in which the dynamics may eventually emerge and impose itself, and one in which the agent can actively engage in trying to achieve a desired ec-pattern.

Third, dynamics like those expressed in t1 are not to be taken as sequences of ordered static contextual episodes. We may be lead to think like that given the way we linguistically encode t1 as a set of clear-cut variables. But this is misleading. Rather, what's represented is the trajectory of the brain's patterns of connectivity. Encoding a trajectory in linguistic terms such as "from p1 to p2" leaves out a lot of potentially important processing that happens as the context switching takes its time. As discussed in chapter 2, the possibility of depicting the dynamics of a process like this is an important purchase of structural schemes that is beyond the reach of schemes carrying absolute contents.

[17] Evidently, the numbers do not mark the sequence in which the patterns happen every day, otherwise they would be sequential. The point of the numeric identification is just to show that the dynamics of a routine may comprise patterns of connectivity that were first realized in very different moments of the agent's life.

The templates we live by resemble — with varied accuracy — the templates comprising the world of human activities. No one alive participated in the invention of the "working 9 to 5" routine, and yet a lot of our inner dynamics resembles that. A similar reasoning works for the brushing-teeth routine just depicted. Given this, it is tempting to paint a picture in which, as we learn our way in the world, culturally inherited routines, habits or traditions yields us all we need in order to store information about the world in the appropriate way. Other people teach us, providing both feedback and encouragement. They point towards the right learning targets: what to do, when and how. Such targets constitute the norms against which our learning attempts must be contrasted with. If we're not there yet, error signals such as "no, this is how you do it" - or non-verbal gestures of disapproval with similar meaning - makes us notice the mistake, and we can keep trying. Thus, the dynamics of the many ec-patterns is inevitably isomorphic to the way that the correspondent activities are articulated in the world. The situations in which you learn about something are the situations in which that knowledge is going to be used.

The reason why this picture is tempting is that it appears to avoid worries about circularity. Lots of local culturally-pervaded learning tasks give rise to the culturally adequate way to store that information. It gives you what's linked to what, and how strong that link is. Thus, it might look like whenever the system needs to stay tuned to what's relevant, all it has to do is follow the way information was organized. As we engage in local learning tasks, we inevitably store them according to the structure of the world. The long-time goal of finding the right way to organize knowledge can thus be achieved in a blind and emerging way, at least from the system's perspective.[18]

Tempting as it is, I think this picture must be rejected. The problem is that social learning is frequently non-local. Whenever someone provides instructions, learning targets or maybe problem formulations in the hope that we can solve them, one expects us to rely on a proper articulation of what is already known. This frequently implies the exploitation of what we're already familiar with in new and creative ways. In other words, human social learning often relies on commonsense, which implies being sensitive to what's relevant. This is true even of kids after a certain age. We certainly don't expect them to realize things before they're ontogenically capable to do so, but we don't give them commonsense classes either. Mostly, we expect them to exploit what they know as we present the inputs, as well as provide encouragement and feedback throughout their developmental trajectory.[19]

The upshot is that even what looks like local learning processes may frequently rely on the complex context-sensitive articulation of what we already know about the world.[20] This

[18] This is the picture adopted by Clark (2002), which I partially reject for the reasons about to come.

[19] Some might complain that, in a way, we do teach them commonsense as we instruct them on what they can and cannot do. True enough, but that's not the point. What's at stake is that we don't need to provide full-fledged instructions of everything we expect from them, as we do with computers. In normal circumstances, "make sure your little brother don't get hurt while playing" is clear enough.

[20] I suspect it is no coincidence that, at a certain point of childhood development, providing guidance feels less like feeding information and more like helping her in properly articulating what she already knows.

is why merely pointing out that the way we learn resembles the way the world is, falls short of a complete answer. The way we store things is not the sole product of many local learning processes. A large amount of integrative capacities must be already operative. Consequently, relevance-sensitivity must be part of the storing-information tale from the start, or at least it must enter the stage very soon. This means that there has to be more involved in what allows us to go beyond the mere familiarity with our previous experiences. The resemblance between the organization of the world and the organization of the stored information provides just a partial answer.

4.5.1 Storing relevance-sensitive knowledge about the world

As information starts to get stored within the cognitive apparatus, it is indeed structured in a way that partially resembles the structure of human activities in the world. For instance, the things we learn first are linked first. And as we learn or practice more, we not only reinforce those links, we create new patterns of linkage between what we're learning with what is already known. But how can all of this knowledge (or some, depending on our current integrative capacities) can bear on and constrain new episodes of learning?

The first step is to notice that, given the aforementioned structural resemblance, Cummins' theory of representation invites us to think that the available structure of such patterns as effectively representing the structure of human activities (at least the portion with which we already engaged with at any moment of our developmental trajectory). But not just that. It also represents the links among the different available portions of stored information and their peculiar dynamics. In other words, it represents the dynamics of the transition from one ec-pattern to another, that is, it represents the system's dynamics in the sense previously outlined (like t1).

Such structure does not depict the full-fledged contents of the cognitive assets involved in the particular functionally identifiable mechanisms articulated, but it does resemble how they were articulated throughout our developmental trajectory. In a nutshell, the system's dynamics comprise a second-order structure that captures how the information is routed and exploited brain-wise (i.e. inasmuch as the brain's physical architecture allows) This kind of representational content is available for exploitation and redescription with the whole set of tools described and discussed in earlier chapters. Thus, rather than taking these patterns to comprise solely resemblances of situations within the human world (being in a restaurant, being in a classroom, and so on), we can concentrate on the dynamics of the ec-patterns themselves. For instance, daily, weekly or even yearly dynamics can be represented and exploited in this way. Such dynamics comprise a second-order structure that captures how the information is routed and exploited brain-wise.

We have thus a natural source of information about what the cognitive apparatus is doing. And we have some plausible candidates for the type of articulation that can be done with it. By capturing the way information travels and flows brain-wise, one is capturing crucial aspects of its pattern of effective connectivity, i.e. its dynamics. This is precisely the kind of structural information that approaches like that of Ribeiro; Saverese; Figueiredo (2017), which we discussed in chapter 2 are well suited to capture. Thus, redescription processes employing this kind of trick may be able to find roles like that of being an authoritative neuronal subpopulation. Authoritativeness, remember, is not identified by the role that a mechanism plays within a given situation, but rather by the relationship it bears with every other mechanism it is able to integrate with - for instance, the “teeth-brushing pattern” tends to impose itself, and bring its comprising mechanisms along with it, every cycle of x hours.

But who’s responsible for doing this work within the brain? We don’t want to go back to the image of a central processor representing and exploiting the dynamics of other processes so it can modulate all subsequent activity. What’s needed is a plausible sketch of how distributed sets of distinct functionally individuated mechanisms could participate in the story. We already have a place to look for the answer. Earlier we saw that neural control structures can route and modulate the flow of information in a context-sensitive manner without implying the need for anything like a central unit for wide-scale processing. Neural control structures share a function, not a place. Unfortunately, as per van Essen’s previously presented statements, in order to do that they relied on innate information about the specific domains they specialized in. Nonetheless, in the same work Essen; Anderson; Olshausen (1994) already speculated that a small step would be enough to render the idea of neural control structures applicable to higher level cognition.

> It requires only a modest conceptual leap to suppose that analogous routing strategies may be used to control the flow of information in whatever central structures are used to represent semantic information and other high-level abstractions that are the coinage of cognitive function. (Essen; Anderson; Olshausen, 1994, p. 298)

Luckily, representational redescription purchases a way to render van Essen’s speculation plausible. A possible story is one in which a mechanism can learn to become a neural control structure and modulate the flow of the mechanisms it happens to interact and integrate with (let us call this set of mechanisms its encompassing subsystem). The candidate to neural control structure is able to learn and representationally redescribe the ec-patterns As it points its learning abilities towards a bunch of mechanisms, it concentrates not on the knowledge that comprises then, but in their usage-dynamic. In other words, rather than capturing information about the knowledge stored in those mechanisms, it captures their dynamics over time, perhaps at multiple scales: when they’re active and who they’re interacting with. In doing so, it represents the informational flow that’s employed and become subject to be further exploited by itself and other eventual consumers in order to modulate that flow, just like it happens with innate neural control structures. Such mechanisms can thus actively participate in every cognitive task involving mechanisms within its subsystem.

Together, these subsystems help to create the overall brain-wide pattern of effective connectivity in which every task is going to be embedded. Thus, at this level of analysis, we can see that there is no central, brain-wide control structure capturing and modulating the informational flow in the cognitive apparatus. Rather, there are many distributed ones capturing distinct aspects (handled by distinct subsystems) of the structures comprising the brain's dynamics. Ultimately, this means that, talking about brain dynamics in terms like sets of patterns like t1 is an obvious simplification. There's not a single global mechanism hiccuping signs that it's time to enter p34. Instead, what we have is a bunch of subsystems converging on a complex brain-wide pattern of effective connectivity that theorists will dub p34 for methodological reasons.

This is the answer to the first question. Whenever we learn our way in the world, we not only store knowledge in the form of a certain structure. Thanks to neural control structures, we store knowledge about the dynamics under which the knowledge is historically articulated as well. This knowledge is implicit in the usage-pattern of the involved cognitive assets and is equally available for further exploitation. We can say that this knowledge resembles what's culturally relevant because it is a direct product of the way in which the world's dynamics routinely take place. Furthermore, at this point, many culturally inherited framing functions can already be involved in the making up of the wordly knowledge. I'm not emphasizing this point yet because, until now, their role is not peculiarly that of comprising cognitive gadgets. That is, up to this point, the story would still be compatible with a set of cognitive instincts exploiting cultural information. But this will change soon.

4.5.2 Relevance-sensitive influence over cognitive processing

The second question is how this knowledge that represents — with varied accuracy — what is culturally relevant can affect cognitive processes. We've seen that, through the lens of RCP, neural control structures can represent and learn to exploit the dynamics of the subsystems they comprise. Whenever they engage in the representational redescription of the dynamics of a subsystem, the output is the outline of a new emerging domain akin to the ones discussed in chapter 3. This means that some of what's typical or cyclical in that dynamics can be representationally redescribed into architectural constraints. As the subsystem's developmental trajectory goes on — i.e. as it enriches its dynamic patterns — it increases the accuracy with which it represents the dynamics of the human activities it participates in. They can effectively learn to render permutations of its previous experiences (inasmuch as its encompassing subsystem participated, of course).

As the subsystem is operative, all that information is available for exploitation. This means that its neural control structures are not restricted to merely reenacting previous dynamics. Rather, the system can effectively exploit simulations of the architecturally available set of permutations of the dynamics it represents ("what if I have lunch earlier today?"). It can forge and "experiment with" such permutations in a flexible yet architecturally constrained and FP-free way. Furthermore, such patterns can be cognized by other mechanisms through additional representations or encodings. Distinct mechanisms can, for instance, produce analogies, compare schemes, task-embed things in it, and so on. In this picture, neural control structures are a source of guidance for other cognitive processes. They provide reliable knowledge of what is typical. However, they can go beyond that. This extra degree of freedom is what allows the system to go beyond what's merely familiar.

Unfortunately, inasmuch as we go beyond what we're familiar with (e.g. non previously experienced permutations of ec-patterns), they can't guarantee that this will happen without losing sight of what's typically relevant. More precisely, once neural control structures redescribe the set of dynamics into a representational scheme and start exploiting it, they become capable of modulating the flow in novel ways. Inasmuch as they have this freedom, they can no longer guarantee that every permutation is going to happen in a way that respects what's circumstantially relevant. After all, being able to render new permutations implies the possibility of moving away from what's typically relevant. How can it preclude contextually alien permutations?

This is where Heyes' framework starts to pay dividends. As we learn our way, we gain access to exploitative strategies culturally established, i.e. cognitive gadgets. The gadgets in question are framing functions that we employ when exploiting some cognitive assets. Such framing functions are akin to the tweaks employed to know whether elephant skin is essential to elephant recognition. In the particular case of neural control structures, the framing functions can tweak our way towards the right direction whenever we exploit possible permutations of our cognitive dynamics. As an example of this point, suppose that whenever doubt obtains, our routine dynamics tell us to follow the highest authority available, and that this usually boils down to following the oldest person in the room. Though this seems like a plausible strategy, as Heyes remarked, this can be misleading, at least in circumstances involving (very) recently developed technology. In these cases, younger people are frequently better authorities on what to do. "Follow the younger" is the kind of gadget that guides us in how we should exploit the space of possible permutations allowed by some domain-comprising dynamics. But the same reasoning goes for lots of very specific outlines: "Hollywood scary movies always try to let you anxious by rising music as the protagonist checks the source of a noise in a dark place, then they suddenly stop the music and let you relax a bit by showing a cat as a possible source, and only then, at that moment of relief, the threat shows itself". Finally, some tweaks may comprise preparatory rituals or catchphrases like "always on the lookout!". Notice how comedy movies typically draws on our familiarity with this kind of framing functions by twisting the typical pattern in which they're applied (things such as a goofy character screaming "always on the lookout!" in a very unusual setting). Thus, lots of small culturally inherited gadgets like these comprise tools that the system can use to guide its exploitation of the possible articulations of its ec-patterns.

This supplies the system with the required flexibility without begging questions about how they can be harmonic with what's culturally relevant, and with principled means of control. It enables us to go beyond what's merely previously familiar by exploiting permutations of ec-patterns in a productive yet culturally constrained way. In this picture, the capacity to stick to what is relevant does not rely solely on how the information is organized and stored within the system. Rather, it is also a function of how the system exploits that information.

The upshot is that what's relevant comes not only in the form of cognitive knowledge about the domains directly involved in human activities. It also comprises knowledge about how to exploit them, i.e. how to process them. This is RCP's way to make the same claim made by Heyes: culture provides not only grist, but also mills, i.e. the means to process grist. This points us to an important clue of how RP can be tamed. It already tells us how we can go beyond what we're familiar with without losing track what's relevant. Whenever we go, i.e. whatever permutation of our typical patterns of effective connectivity we exploit, we'll remain loyal to what's culturally relevant inasmuch as we rely on cognitive gadgets to tweak our exploitation of our own dynamics.

Unfortunately, there's still one last step before reaching a complete account of RP. As things stand, we may still have a potentially overwhelming number of small gadgets like the ones aforementioned. Thus, the possibility arises that we are again haunted by the problem of how to choose the right gadget to apply at any circumstance. In the next session, I'll make a tentative suggestion to tell a story in which this can be avoided.

4.5.3 Relevance-sensitive context extrapolation

Extrapolating from familiar contexts without losing track of what's relevant is the core cognitive capacity underpinning the possibility of handling non-saturable contexts. That amounts to the capacity to engage in novel variations of already familiar contextual tonalities even though we can't know in advance the kind of feature we'll have to handle. We've seen how that can be done with the help of culturally inherited cognitive strategies (i.e. cognitive gadgets). An important part of the story is that we should not concentrate on kinds of situation, but rather in the system's dynamics. Whenever we find ourselves within a situation, we're handling it under some articulation of the possible ec-patterns. The cognitive tweaks we culturally inherited work with them as the background against which they're deployed, not with clear-cut situations. Thus, ec-patterns and situations comprise distinct dimensions of the same structure comprising human activities. This means that we can concentrate on one dimension rather than the other, and that this decision will always leave something out.[21] The current point, to be clear, is not that the dimensions exploited whenever trying to demarcate the boundaries of a concrete situation are useless or not exploited at all. Rather, the point is that they're not behind the system's capacity to keep track of what's typically relevant in indefinitely many contexts.

We can summarize our current goal like this: how do we avoid idle permutations of our ec-patterns? If neural control structures (as well as whatever mechanism exploits them) are allowed to free float, RP will remain a genuine threat. A while ago, Cherniak (1990) reached a similar point. He realized that the way we store information is intertwined with our capacity to

[21] That's expected whenever handling multi-dimensional structures, as extensively discussed in chapter 2.

exploit it in a relevance-sensitive way. Also like us, he noticed the limitations of this approach. Storing in the right way seems necessary, but not sufficient to explain our capacity to render plausible and relevance-sensitive inferences. Unfortunately, at the time, all he did was to speculate that we somehow learn the right way to do that or inherit the proper genetic wiring:

> (...) actions of inferring must (...) come to conform with desires and beliefs largely by means of fixed, nonintegrated, nonconscious mechanisms of selection or guidance that do not involve reasoning processes of any kind. These mechanisms may be acquired—for instance, as learned "cognitive styles" — or natural selection may have "designed" the agent so that, as an efficient organism, he undertakes particular inferences. (Cherniak, 1990, p. 12)

We know why the natural selection of genetic traits is hopeless: our upbringing relies on culturally inherited gadgets. As for what Cherniak dubbed *cognitive style*, unfortunately he didn't develop it, keeping it somewhat mysterious. After all, acquiring a cognitive style is nothing but learning to make the appropriate inferences in a relevance-sensitive way. Moreover, Cherniak thought that a cognitive style must not involve any kind of reasoning processes. The reason for this claim is that Cherniak worked with a LOT-powered framework. In that scenario, any appeal to reasoning processes would be circular, that is, it would beg the question that a style was supposed to solve, which is how we can engage in inferential productivity that respects what's contextually relevant. Notwithstanding, the term "cognitive style" seems appropriate to RCP. That's not because it designates non-cognitive work, though. It's because it designates a collection of strategies to render inferences that do not work with a principled notion of what's relevant. In other words, the way we do things within a given culture is not a feature of some principled conception of relevance.[22] Rather, it is the result of the accumulation of contingencies that characterizes cultural evolution. It is not merely a cognitive routine, for it's what allows one to go beyond what's cognitively routine. But what could a cognitive style amount to? The set of conceptual tools involved in RCP and the cognitive gadgets framework may be a fruitful way to flesh out the idea.

A cognitive style can be regarded as a set of strategies employed by mechanisms exploiting the stored structural knowledge about the dynamics of cognition. However, this time what I have in mind is something different from the small cognitive tweaks mentioned in the previous session. Those tweaks are surely available to be inherited and employed as previously stated, but there's a distinctive kind of cognitive gadget that comes from the cultural accumulation of those small tweaks. Over time, they comprise what I'll regard as *cognitive templates* for the overall dynamics of the cognitive apparatus. As the name emphasizes, such templates comprise the dynamics of ec-patterns, just like the ones previously discussed. However, templates comprising cognitive styles are usually large-scale, in the sense that they can involve the operation of many distinct routing mechanisms. In other words, they're templates for a certain way to render the dynamics of the world intelligible. To get a glimpse on what such

[22] This is just a restatement of an idea already worked out in chapter 1: there can be no general theory of relevance.

strategies would amount to, I'll take an idea from another old paper, this time from Dennett (1996), which is worth citing at length:

> Suppose an agent is supplied with half a dozen story outlines — or 25 or 103, but not many more — and these are so over learned that these stories are always "trying to happen". That is, whatever happens to the agent, the stories are crowding around, looking for their particular heroes, heroines, obstacles, opportunities and circumstances, and very tolerant of mismatches. (...) Suppose, that is, that the difficulties the hero encounters and deals with in each story is typical, and the repertoire of stories is *practically* exhaustive of human problems. Each story, as it "tries" to be reenacted in the agent's own experience, provides the agent with a model of how to behave. It keeps giving the agent hints: Stay on the lookout for dirty tricks of sort A, do not forget to take the key along, perhaps this is an opportunity to try the old "push-the-witch-in-the-oven" trick, wait till the villain falls asleep, perhaps this is a wolf in sheep's clothing, and so on. (...) An agent equipped with a sufficient stock of stories trying to happen would always be vulnerable to being blindsided by an entirely novel story, but perhaps, after millenia of storytelling, there are no more entirely novel stories. That, at least, is the conclusion despairingly reached by many a blocked novelist and triumphantly reached by many a structuralist anthropologist. (Dennett, 1996, p. 6)

Dennett suggests that we have a number of "story outlines" that we employ in trying to make sense of the world. His idea can be easily connected to that of Joseph Campbell's "journey of the hero", from Campbell (1949). In his studies of mythology, Campbell hypothesized that there would be a common structure underlying the myths that human beings have developed throughout their history. Very roughly, the structure would comprise a dynamics in which the protagonist is called to an adventure, faces and wins a decisive challenge and then returns home transformed by the experience. But we're not concerned here with how credible Campbell's thesis is in anthropology. The idea of a "monomyth" was much more ambitious than it is required by the illustrative role it plays here. For instance, neither Dennett nor I have in mind something that necessarily crosses different cultures. Quite the contrary, what matters for us (well, for me) is the possibility of structural dynamics culturally inherited by cultural learning, perhaps sometimes by literally hearing about them in tales.

Many of such story outlines may resemble large-scale expected dynamics in multiple domains of life, such as relationships, health and work. They may also resemble patterns of a shorter time scale, such as those involved in telling stories with a clear "moral" ("be careful, you know the story of the boy who cried wolf"). They can even resemble the behavioral patterns of particular agents, perhaps in a religion-embedded sense ("this is what she would do", "this is what he would like"). We learn the trajectories comprising such stories as we learn to navigate in the world. Consider, for instance, the dynamics of friendship alongside a large-scale temporal dimension. It usually begins with an encounter involving common goals, interests or experiences. There are many steps in it as the level of intimacy grows (or perhaps remains stationary in a plateau). It implies duties and roles, in the sense that it has to be nurtured and leads to expectations. Moreover, you're subject to having favors asked, you're expected to be loyal, to be there for good and bad moments, to remember special occasions, and so on. As we master the whole dynamics, it can play the role that Dennett attributes to it and keep "trying to happen", occasionally tipping us by whispering "this is the part where you're expected to do x".

Dennett's story outlines, I suggest, are the peculiarly human way to generalize about the world. In RCP, they amount to cognitive templates, i.e. to set of cognitive gadgets structured by cultural evolution. In their turn, such templates comprise the cognitive style that resembles what's culturally relevant in a given culture. Cognitive templates enable us to handle non-saturable contexts because they're always operative in whatever concrete situation we find ourselves in. Even in situations that we're never found before, they can point us towards all kinds of cognitive tricks. What to pay attention to, what is likely to come next, and so on. Thus, we can go beyond the set of permutations of the patterns of effective connectivity with which we're already familiar without getting lost: on the one hand, architectural features of the employed representational schemes preempt the possibility of going astray due to FP. On the other, the availability of a cognitive style, i.e. a relatively small set of cognitive templates always "trying to happen" preempts the possibility of going astray in indefinitely many idle permutations. They're available for both learning and non-learning cognitive tasks.

The importance that the templates comprising a cognitive style are limited in number should not be underestimated. In order to avoid RP, we need a good balance between flexibility and stability. The availability of a reasonably limited yet powerful set of templates for stories might be able to provide just the right amount of each. A huge number of cognitive templates would simply bring RP back, for now we would have the problem of choosing among them. On the other hand, the absence of templates would keep us in the cage of the worldly experiences we're previously familiar with.

This implies that the human-world comprises also templates rather than just kinds of situations.[23] In this view, as stated in the first chapter, a cognitive context does not amount to a subset of the cognitive apparatus' cognitive assets, but rather a tonality. Thus, even those willing to tell their tales by relying on m-contexts need not worry about what is in or out a specific situation. Everything is potentially in, but the system explores the possible configurations only in terms of how it articulates the permutations of ec-patterns allowed by its cognitive style. Its cognitive style is, in this sense, the set of possible ways to articulate the available ec-patterns, both acquired as the system learns about the world through culturally inherited framing functions. This is how the system is able to keep its interests and the world in harmony. The system's current contextual tonality is the system's current articulation of the available ec-patterns alongside the world's circumstances. As the set of available permutations of ec-patterns can be articulated productively without raising RP, it is not mysterious how the system can entertain an open-ended set of non-saturable contexts.

Before we dig further into how this enables a system to switch from context to context without raising RP, two quick remarks are in place. First, in her book, Cecilia Heyes

[23] Indeed, there's the possibility that situations only make sense from the perspective of the theorist. This amounts to the denial of any explanatory role for m-contexts.

distinguishes cultural learning as a special kind of social learning that's specialized for cultural evolution. I've just claimed that relevance sensitivity is a cognitive gadget, and that it plays a special role in many of our cognitive tasks, both those involved in learning and those not involved. Should we say that relevance sensitivity (i.e. a cognitive style) is specialized for cultural inheritance? I'm inclined to think so. It provides a means of enhancing the fidelity and accuracy with which information is passed on. Furthermore, it helps to gather very similar understanding from and enormous variety of situations. While this obviously requires further research, I think there are good reasons to be cautiously optimistic about it.

The second remark is about whether we should think of relevance sensitivity as a metacognitive mechanism. It all depends on what one means by that. But again, I'm inclined to think so, at least if we accept the conception advanced by Proust (2013), which recognizes a place for what she calls "procedural meta-cognition". In this picture, such mechanisms need not be described according to linguistic or "conceptual" rules. Rather, they may provide clues about what we should regard as relevant in a given situation even though it's underlying rationale is not available to us. A culturally established cognitive style seems fit for this role.

4.5.4 Relevance-sensitive context switching

In RCP, Dennett's story outlines are cognitive templates that can be exploited in order to guide and constrain other cognitive processing. That is, by "adopting" a given template, what one is doing is entering a certain state characterized by some pattern of effective connectivity. Such an articulation amounts to the system's current contextual tonality. But how can a given permutation of a story come about? In other words, how does context switching take place? The usual way to think about this is in terms of the system being able to track the world and render harmonic internal states. This doesn't throw much light on the role played by the system's interests and expectations, though. In what follows I'll quickly describe two possibilities that are likely to coexist: the first is a bottom-up approach that describes the familiar set in which the system simply finds itself within a given context that comes about. The second amounts to a top-down approach in which the system can actively engage into bringing about a contextual tonality.

Perkins (2002) claims that behavior is largely controlled by a bottom-up activity switching mechanism he dubs *emergent activity switching*. This is a special case of *self-organizing criticality*, a concept that comes from physics. It describes the dynamics of systems tending towards an atractor. Perhaps the simplest example is thirst. As the intensity increases, it enhances the likelihood that the system shifts towards water-seeking behavior. There's a critical point, though, where the behavior becomes almost inevitable, i.e. it becomes the system's focus. The criticality is said to be "self-organizing" because it comes about as part of the way the subsystem works. It is not really a perturbation on some stabilized flow, but rather a consequence of what keeps the system (mostly) stable. The closer the thirst-modulating system is to a critical point, the more likely it is to make the system pay attention to external features such as the availability of a bottle of water. Eventually, if the thirst continue to build it may trigger an active attempt to find water. After finally drinking some water, the system enters a “dormant” state in which bottles of water are less likely to be salient.

The critical phase primes the system towards a certain contextual tonality. In RCP terms, this means it makes it enter a sequence of ec-patterns as described by a template. Not all templates must be assembled by culturally inherited gadgets or elements, of course. A thirst-modulating system, if there really is one, is likely to be accounted for solely in terms of genetic evolution. But Perkins’ engine works with culturally assembled templates as well. It invites us to think of context switching as emerging out of many concurrent processes of self-organizing criticality.

This is plausible account of how there can be stories that are always “trying to happen”. Different processes tend to build up to the point where some critical phase triggers a contextual tonality.

> Emergent activity switching appears to be a highly adaptive way of resolving what behaviors to execute when. Forget for a moment that we like to think of ourselves as intelligent organisms capable of managing our behavior in global, planful top-down ways. Perhaps I am not so mindful this week, or sleepy today; or perhaps I am a dog or a cat instead of a human being. Emergent activity switching keeps me drinking when I need to drink, without reliance on self-knowledge and panoramic planning. (…) I don’t have to monitor deliberately whether I’m thirsty. My body tells me. (Perkins, 2002, p. 70)

The picture that Perkins draws in this quote is recognizable in much of the literature regarding RP. Indeed, Dreyfus’ famous skepticism towards classic cognitivism was largely grounded in a description of the background (in which every cognitive task takes place) as a phenomenon that cannot be cognitively explained (Dreyfus, 1992; Dreyfus; Dreyfus, 1987). This pictures the system’s cognitive machinery as a dupe of such low-level undergoings, as if the background in which the system’s cognitive tasks take place was always a given, at least from the perspective of the system’s cognitive activity. In this view, the relation between a context and the tasks that take place in it is unidirectional. This goes both for low level water-seeking routines and routines involving higher cognitive capacities, even when consciously exercised.

> (…) the pattern also figures prominently in loftier activities. Consider, for example, preparing for an upcoming interview. For a while, you don’t worry about it at all. As the time approaches, it’s more and more on your mind. Eventually, you worry enough about it so that you sit down and begin to prepare. The period of preparation continues until you feel secure. Having rendered the worry dormant, you switch out of the preparation behavior, proceeding with other activities until the interview. Or, alternatively, worry may begin to build again as you think of real or imagined shortfalls in your initial preparation. (Perkins, 2002, p. 69)

This brings forth a picture in which we don’t have to monitor deliberately whether it’s time to prepare for an interview (or a presentation, or a thesis defense). We can instead "trust our worry". This is how our cognitive templates interact with and preempt the cognitive tasks we put forth. It brings to the fore certain templates according to certain patterns which are describable as emergent. Inasmuch as RP is concerned, the fact that templates stories are assembled with the help of cognitive gadgets (i.e. culturally inherited tweaks) preempts the necessity to invoke a pre-established harmony between cognitive processes and the structure of worldly activities.

But there's something unsatisfactory with this unidirectional account of the relation between contextual tonality and the cognitive tasks we engage in. Up to this point, this view pictures the system as a dupe of its own cognitive style. Though this is undoubtedly true, it should be emphasized that there is another route towards establishing a distinct context for the realization of a task, i.e. context switching. There are cases in which a system (like us) can actively engage in bringing the proper contextual tonality to the fore. By doing so, we can actively change to a different articulation of ec-patterns. That is, rather than merely attuning ourselves to the current context of activity, we can actually refer to another, non-actual context and change the way we handle a situation. Reconfiguring the context itself may become the cognitive task.

In order to make this suggestion, I'll deploy an example from Cappuccio; Wheeler (2012): how United Kingdom's King George VI dealt with his famous stutter. More specifically, how he managed to handle it when he had to announce that the United Kingdom was entering World War II. Until a few weeks earlier, George was not the king, he was unprepared to be crowned, and thought of himself as a marine officer. His circumstances were really intimidating. The set of contexts with which it was already familiar with are very dissimilar from the one he was facing now. Given his complete lack of familiarity, the disparity between how he'd see himself and his current role, we could hardly claim that his set of familiar templates would poise him to react appropriately. In order to handle it, George made use of preparatory routines such as practicing with tongue twisters and training with preparatory gestures. In the words of Cappuccio; Wheeler (2012):

> It seems difficult to claim that his decision to confront the local context of that public announcement was fluidly coupled with, and unproblematically derived from his personal and political history (...), as the emotional tone of the situation seemed strongly to discourage that decision; so much so, in fact, that even his personal identity and institutional function, and not just the contingent circumstances of the speech, could be seen as an anguishing state of affairs to be avoided. (...) This shift could be achieved only through a resolute effort on the part of the king to become aware of, and to change, the conditions of his contextual situation, and not through a process in which he unreflectively accommodated himself to those conditions. (Cappuccio; Wheeler, 2012, pp. 23–24)

By training with preparatory routines, George could "reset" its behavioral dispositions. He could then bring to the fore a new contextual tonality, and with it, he brought up a distinct articulation of worldly knowledge and behavioral dispositions he could use to act differently in a very unfamiliar context. This requires some level of detachment from the current environmental circumstances, which means that there is a place for the kind of job that representational contents can be specially helpful. Representational articulations of templates are able to handle this, for they can amount to the partial simulation of a distinct contextual tonality. George was trying to employ framing functions to simulate a new permutation of the contextual tonalities he was already familiar with. Such simulation can help him to behave differently in a scenario he has never found before. In order to do so, he used preparatory rituals (such as tongue twisters) that helped him to get closer to the desired contextual articulation. These are culturally inherited scaffolds that enable the system to exploit its set of represented templates in new ways, i.e. it can learn to articulate new templates or learn to apply a given template differently.

The described process should not be taken as a full inner power that can fully overcome external circumstances or bring about a new contextual configuration out of the blue. Rather, we're pointing to a process of context-switching that is not a direct response to environmental changing, but a strategy actively adopted by the agent. The system is able to think about its own dispositions from the simulated context that arises out of the simulated template. By manipulating and simulating a new context, he can "check" what is seen by him as relevant, that is, he can let itself be guided by the articulated template. This is an example of how we can tell a tale in which representations can effectively guide the system into a new contextual tonality without the need to posit m-contexts, but without precluding them through non-empirical means. This allows us to remain faithful to the idea that this should be determined through empirical means.

4.6 Closing remarks

Let us finally outline everything we've discussed and make sure we haven't left out anything essential nor any promises unfulfilled. What have we (hopefully) achieved? Ultimately, relevance-sensitivity is the product of a cognitive gadget dubbed cognitive style. This gadget is produced by cultural evolution and inherited through social learning. As conceived, this notion of relevance-sensitivity is completely compatible with representational mechanisms — and if RCP is anywhere near the right path, representational accounts may provide the best trade off between explanatory purchases and theoretic commitments. It is algo non-algorithmic in the sense that there's no underlying principle or theory of relevance behind what constitutes relevance-sensitivity, i.e. it is compatible with the claim that there can be no such thing as a contextless algorithm to determine what's relevant. Rather, it is compatible with the idea that relevance-sensitivity comprises a broadly pragmatic engagement with the world.

Can a cognitive style handles non-saturable human-world contexts? That is, can it buys us relevance-sensitive context productivity? A cognitive style supplies the system with a finite and limited set of exploitative strategies (i.e. templates). Consequently, there's only a limited capacity to render permutations of the systems' patterns of effective connectivity (ec-patterns). Distinct permutations result in the articulation of different local mechanism in different ways. This means that different kinds of elements will become prominent in different degrees of intensity. As the cognitive templates comprising the system's style are always "trying to happen", their sensitiveness to certain key clues can quickly render a critical point leading to a context switching. Thus, a friendship-template can quickly lead to a health-template in case a friend starts to not feeling well during dinner. But this doesn't stop the friendship-template nor the (say) social-relationship-template from keep "trying to happen" at all times.

Ultimately, there may be limitations to the kinds of feature that can become salient. But this limitation is more the result of the volume of knowledge stored in the system than the scope of the templates it continually employs. That is, a template cannot bring to the fore elements that are completely alien to a particular system. However, as the system learns about those elements in some other occasion, the same template now can render that element salient whenever the involved mechanism gets active under a given pattern of effective connectivity. Thus, eventual limitations are not cognitive, but epistemic. They're not connected to the depiction of contexts as kinds of situations. They're indifferent to what's typical to a given circumstance, for the specifics of each situation is only indirectly connected to the templates. In other words, if a cognitive template involves a given articulation of 24 different mechanisms, and none of them involves any sensitiveness to a feature x, then no permutation of those mechanisms will enable the system to render that feature x as salient. But by the same token, if any of those 24 mechanisms eventually learn about the feature x, then the same template will be able to render it salient in any situation. That is, we can become attentive to an increasingly higher number of clues as the comprising mechanisms learn something new, for the sensitiveness is now part of every pattern of effective connectivity rendered by a given template. We have, in a nutshell, a productive yet tamed capacity for context production that can extrapolate from the set of familiar contexts without implying that the involved cognitive mechanisms are able to spin freely in a void.

Finally, RP was first introduced as a challenge facing accounts of two large-scale human capacities: commonsense holism and situation holism. Can the provided account handle them inasmuch as RP is involved? Time will tell, but I'm cautiously optimistic. Commonsense is hard inasmuch as it relies on open-ended sets of non-saturable contexts. But our cognitive style provides context-productivity, that is, as we go from context to context, our templates guides into what must be rendered salient, whatever that context amounts to. Moreover, the fact that the cognitive templates are always operative helps us to explain the real time character of commonsense. It accounts for the fact that we continuously reassess information as it flows. In this sense, to manifest commonsense is simply to constrain one's own inferential productivity to one's cognitive style. It places our cognition between the Scylla of an inflexible routine and the Charybdis of a completely unconstrained inferential process.

At last, the capacity to keep track of two (or maybe more) situations at the same time is not mysterious anymore. The context productivity afforded by our cognitive style, as well as the bottom-up and top-down mechanisms of context-switching can be articulated in parallel without begging any fundamental question. There can be of course empirical concerns about the practical limits of such capacities. How many distinct perspectives of the world are we capable of articulating at once? Nothing in RCP puts a principled limit to this capacity.

4.7 Paragraph-by-paragraph summary

Each entry below summarizes a paragraph of the main text.

Where are we at?

Let's start by taking stock: RP seems to bother everyone.

But not equally, for those insisting on systems employing absolute contents face a dead-end, which is FP.

Given everything we discussed, it seems clear that determining whether something is relevant cannot be the goal of any particular, cognitive task, for it would suffer from RP itself, which means that relevance must be some kind of emergent property. This is problematic, though.

On the one hand, the claim that relevance-sensitivity must be a consequence of the system's overall dynamics is specially worrisome for those commited to the idea that cognition can only be properly explained by subsuming performances under some kind of epistemic rationale.

On the other, there seems to be no underlying rationale capable of figuring out what's relevant.

Switching from absolute to relative contents helps by getting us rid of FP, but it won't do much more than that RP-wise.

This is so because every representational content (as well as every target the system fixes on) is aspectual.

Going for a non-cognitive account is not an option, for relevance sensitivity must be potentially exploitable as a source of knowledge, otherwise we wouldn't be able to extrapolate from it, and extrapolation is essential to handle non-saturable contexts.

My suggestion for achieving this will rely on structural representational contents and the capacity for representational redescription.

First, we'll discuss how a broadly mechanistic approach can handle the high level of flexibility that context-sensitivity requires, and then proceed to the suggestion of a bootstrap story in which we can "learn to learn" what's relevant with the help of culturally inherited gadgets.

Effective connectivity

How can the brain be flexible yet organized?

There are frameworks that seem to enable a high level of cognitive flexibility, but with a catch: whenever the requirement for flexibility becomes too complex or too broad, only a broadly dynamic account can be formulated (i.e. we give up on a mechanistic account).

But most of the skepticism towards mechanistic accounts of very flexible systems involves conceptions of functional analysis and computational architectures coming from classic cognitivism.

The main issue is that cognitive machinery must be broadly integrated while, at the same time, not restricted to a single organization. But how can such wide reorganization takes place if the reorganizing process must itself rely on the current organization?

The first step is to realize that functional and mechanistic analysis can themselves be context-sensitive.

The second step is to sketch a mechanism through which wide functional reorganization could be achieved: we find a clue in what Andy Clark dubbed neural control structures.

They are responsible to modulate the flow of activity between cortical areas.

The hypothesis is that there are mechanisms specialized in modulating the flow of information, and that these have a crucial role in how distinct mechanisms are articulated and how they can influence one another.

The influence that neural mechanisms have over one another at a given moment comprises the systems' overall pattern of *effective connectivity*, enabling the system to comprise transient sub-systems that are tailor-made for the situation at hand.

But how do neural control structures "know" the pattern of effective connectivity that accounts for what's relevant in the current situation?

Perhaps it can rely on the set of situations (i.e. previously employed patterns of effective connectivity) it is already familiar with and simply bring to the fore the most similar one?

Unfortunately, similarity-based recognition requires (rather than provides) a solution for RP, for the comparison must regard only the relevant set of potentially similar features (given that everything is similar and dissimilar in indefinitely many ways). We have thus the required flexibility, but it remains blind to relevance. Perhaps it can learn to see it?

Can we learn what's relevant?

Fodor's classic take depicts relevance as an objective property that must be adressed through scientific endeavours.

But that is misleading, for science is about what's true, while relevance is about what's fit to a given context.

As we learn to inhabit the human world, we learn to cope with what's relevant for the activities that take place within, and in this sense, relevance emerges as a structural trait that's out there for us to cognize.

Unfortunately, even these learning tasks suffer from RP, which is why one doesn't simply learn what's relevant: the learning system must be able to select and track the relevant set of features.

A version of this difficulty can be found in artificial neural networks: sometimes it may take an inessential feature as essential and vice versa.

Thus, we can't simply appeal to the availability of the relevant information in the environment, for the question is just how the learning mechanisms can discern relevant from irrelevant features when learning about any given target.

To see what's at stake, consider the example of the mutilated checkerboard to see how a small difference in how the problem is formulated may require a deep insight about what features are relevant.

Can 32 domino pieces cover the 64 squares of a normal checkerboard? Yes.

Now say we remove two diagonally opposite corners of the board: can 31 domino pieces cover the remaining 62 squares? No. In order to solve it now, we have to pay attention to a so far irrelevant feature: the fact that each square has a color (black or white), and the color alternates.

The point is that the checkerboard has indefinitely many properties other than parity, none of them equally helpfull, but we only know that after realizing what's relevant. How?

This amounts to Clark and Thornton's distinction between type-1 and type-2 regularities. Type-1 regularities can be grasped directly, while type-2 require framing the data in the circumstantially appropriate way. Let's unpack this.

A type-1 regularity enables direct grasping. We check each available input-output pair and come up with a probabilistic conclusion (e.g. half the times that 1 is in the input data, the output is also 1, which means that there's a 1/2 chance of getting 1 as the output).

However, there are statistical relationships of this kind that only become available for direct inspection *after* the application of some arbitrary function.

By framing the input set under a given *framing function*, a whole new set of inferences become immediately available by the same kind of statistical inspection previously available with the raw data.

RCP's previously discussed cognitive tools (schemes, encodings, task-embedding, and so on) are all akin to framing functions in this sense: they may render type-2 information to type-1 exploitation.

However, as the authors notice, now we have the problem of selecting the contextually appropriate way to render the available information intelligible among indefinitely many possible ways.

Selecting the contextually appropriate framing function under which to subsume information is essential, and it is also what triggers RP even in regular episodes of learning. This is why one can't simply learn what's relevant.

In learning, a framing function is akin to a developmental bias, but where do these biases come from? Let's consider some possibilities.

Option one: general-purpose learning

The first possibility is that general-purpose learning are able to figure out type-2 regularities without any framing functions.

This idea is reminiscent of the hope that we can find *the* algorithm comprising general intelligence.

Though some non-specialized algorithms can figure out some type-2 regulatities, this is not fully generalizable.

As the "no free lunch" theorems have shown, every learning algorithm will inevitably embody certain domain-specific assumptions that will favor some framing functions over others.

Option two: meta-learning

Meta-learners try to learn the best learning approach for each circumstance, but the threat of infinite regress is evident.

Vervaeke and others believe that this threat can be mitigated if we conceive the meta-learner as relying on opponent processing (i.e. it would learn the best balance between competing goals).

Every subsystem would be sensitive to specific elements of the environment, somewhat like the resulting balance of the sympathetic and parasympathetic systems. The meta-learner would work to achieve the appropriate balance in every situation.

Thus, finding what's relevant is understood as the capacity to find the contextually appropriate balance between the many opponent processes going on.

This is a nice step in a promising direction, for it takes relevance to be not just about efficiency, but the balance between efficiency (using one's cognitive knowledge in the best possible way) and diversification, which the authors dub resiliency (trying out new things, even if it may lose time and energy with inneficient attempts).

But this is just another kind of CRC (which we studied when discussing Wheeler's approach in chapter 1) and thus subject to the same kind of criticism: it may account for the cognitive flexibility, but it throws no light on what guides the system in figurint out the contextually appropriate balance.

What we have here, I think, is another attempt to stop a vicious regress of reformulations disguised as explanations of RP by appeal to a change in the explanatory strategy.

Option three: innate tweaks

A different approach involves heavy reliance on innate framing functions: we're born with the appropriate set of tweaks programmed to come to the fore at the right developmental stage.

Though plausible in some cases, this leaves a lot unexplained: we most certainly do not genetically inherit developmental tendencies to learn the right way to behave in classrooms or airplanes.

Option four: incremental learning

The idea behind incremental learning is that complex type-2 problems can be factored in developmental sequences of easier type-1 problems.

Unfortunately, as it stands this is just a reformulation of the problem: how do we figure out the right developmental sequence?

Either we have a set of framing functions bounded to a specific domain, or we have a set of framing functions that tries to handle every domain equally, without the proper treatment of its particularities.

This difficulty is similar in nature with those we've been discussing so far. We can't simply learn what's relevant.

But I believe we can find a way out of this maze in Cecilia Heye's framework of *cognitive gadgets*.

The cognitive gadgets framework

As things stand, we already know that in a very general and explanatorily trivial sense, we learn to recognize what's relevant as we learn to cope with the world. However, as we try to make sense of this capacity at the sub-personal level, things remain quite challenging. Perhaps Cecilia Heyes' framework can be of help.

Heyes goes beyond the traditional picture (in which we use innate mechanisms to process cultural information) by claiming that we can culturally inherit also the set of techniques we employ when handling cultural information.

To ground this, Heyes provides a selectionist account of cultural evolution: mental mechanisms evolve and are replicated through social learning.

Social learning can do that because it's not only about "copying" information from a mind to another, but rather gradually shaping one's behavior through continuous instruction and encouragement.

Thus, rather than concentrating on finding peculiarly human "cognitive instincts", we can instead focus on accounting for what's peculiar about human culture.

Heyes ground's what's peculiar in our culture in three genetic traits: higher sociability, enhanced social motivation and more powerfull general-purpose information-processing machinery.

Thus, the distinctive innate developmental biases are those that enhance our sociability.

These traits enable a kind of social learning that's responsible for the peculair traits of human culture. Heyes calls them mechanisms for cultural learning, which are specialized for our peculiar cumulative culture.

In this picture, the claim that cognitive gadgets are products of cultural evolution amounts to the claim that they were shaped by cultural group selection.

An example of how a cognitive gadget can develop out of the presented starter pack is the human capacity to imitate.

The problem, in a nutshell, is how the cognitive apparatus can associate the movement the agent produces with the movement from third-parties that the agent only sees.

Her core idea is that we learn in increasingly larger repertoire of associations between what we see and what we do: babies can learn such associations by self-observation and by being exposed to their mother's imitation of their own behavior (frowning, etc.).

This kind of association is an example of what framing functions can look like in the real world.

Only now, thanks to Heyes' framework, we have a plausible source for culturally shaped framing functions and their role in developmental trajectories.

This is a cultural account of why humans are so good at imitation: even though we may share some employed capacities with other primates, their environment is not so rich (neither cumulative) as the one available for humans.

Culturally inherited framing functions can explain developmental trajectories that only make sense within the current setting in human culture.

In a nutshell, by providing the right set of framing functions at the right developmental stage, culture is able to provide the right developmental sequence of tractable type-1 learning problems.

Relevance sensitivity as a cognitive gadget

Let us now concentrate on how what's culturally relevant gets stored and how it can constrain further cognitive processing.

Every contextual tonality involves a correspondent pattern of effective connectivity (ec-pattern).

The dynamics comprising a sequence of ec-patterns may constitute routines, and such routines may be deemed as templates for the system's dynamics.

But routines are not the only source of templates, for they can be learned through social learning as well.

Moreover, at any given moment, there can be multiple templates taking care of their business in parallel.

Also, it is important to remark that a template depicts a trajectory from one ec-pattern to another, so there's no clear-cut boundary between them.

The templates we live by resemble — with varied accuracy — the templates comprising the world of human activities, so it is tempting to claim that, as we learn our way in the world, culturally inherited routines, habits or traditions are all we need to store information in appropriate ways.

This is tempting because it avoids worries about circularity: lots of "local" culturally-pervaded learning tasks give rise to the culturally adequate way to store that information. The long-time goal of finding the right way to organize knowledge can thus be achieved in a blind and emerging way, at least from the system's perspective.

But we must reject it because social learning is frequently non-local. Whenever someone provides instructions, learning targets or maybe problem formulations in the hope that we can solve them, one frequently expects us to rely on a proper (and sometimes creative) articulation of what is already known, i.e. one relies on our commonsense.

The upshot is that even what looks like local learning processes may frequently rely on the complex context-sensitive articulation of what we already know about the world. Consequently, relevance-sensitivity must be part of the storing-information tale and there has to be more than a mere resemblance between the organization of the stored information and the world. We need to dig deeper.

Storing relevance-sensitive knowledge about the world

At first, some information is indeed isomorphic to what is out there.

This results in structural representations of the world's dynamics that are available for exploitation.

Such structure does not depict the full-fledged contents of the cognitive assets involved in the particular functionally identifiable mechanisms articulated, but it does resemble how they were articulated throughout our developmental trajectory. Such dynamics comprise a second-order structure that captures how the information is routed and exploited.

These ec-patterns can be representationally redescribed into full-fledged representational schemes.

But who's responsible for doing this work within the brain? We have to avoid getting back to the image of a central processor.

Multiple mechanisms can learn to become a neural control structure and modulate the flow of the mechanisms it happens to interact and integrate with (i.e. their encompassing subsystems): they don't capture information about the knowledge being employed, but their dynamics over time (when they're active, and who they're interacting with).

Together, these subsystems help to create the overall brain-wide pattern of effective connectivity in which every task is going to be embedded.

This is the answer to the first question: the system does not store only knowledge in a way that resembles wordly activities, it stores information about the system's own dynamics as well.

Relevance-sensitive influence over cognitive processing

The second question is how this knowledge that represents — with varied accuracy — what is culturally relevant can affect cognitive processes.

As neural control structures representationally redescribe the system's dynamics, they can effectively exploit simulations of the architecturally available set of permutations afforded by the underlying scheme, and they can provide guidance for other tasks. This allows the system to go beyond what's merely familiar.

But how can the system preclude an explosion of contextually alien permutations?

Again, Heyes' framework can be of help: lots of small culturally inherited gadgets comprise tools that the system can use to guide its exploitation of the possible articulations of its own dynamics.

This supplies the system with the required flexibility without begging questions about how they can be harmonic with what's culturally relevant, and with principled means of control.

Thus, relevance-sensitivity involves culturally inherited knowledge about how to exploit cultural information. This way, whatever permutation of our typical patterns of effective connectivity we exploit, we'll remain loyal to what's culturally relevant inasmuch as we rely on cognitive gadgets to guide the exploitation of our own dynamics.

One threat remains, though: there may be an overwhelming number of gadgets to employ, and we need to abide by the relevant ones. How?

Relevance-sensitive context extrapolation

Extrapolating from familiar contexts without losing track of what's relevant is the core cognitive capacity underpinning the possibility of handling non-saturable contexts.

Thus, we can summarize our current goal like this: how do we avoid idle permutations when exploiting representations of our dynamics of ec-patterns?

What we need is what I'll dub cognitive style: a culturally established collection of strategies used to make inferences without loosing track of what's culturally relevant. A cognitive gadget that embodies relevance-sensitivity.

A cognitive style is constituted of cognitive templates for the overall dynamics of the cognitive apparatus: these are large-scale, in the sense they can involve many distinct routing mechanisms.

Such cognitive templates bois down to what Dennett called “story outlines” that humans employ in trying to make sense of the world.

Many of such story outlines may resemble large-scale expected dynamics in multiple domains of life (at multiple temporal scales), such as relationships, health and work. As we master the whole dynamics, it can play the role that Dennett attributes to it and keep “trying to happen”, occasionally tipping us by whispering “this is the part where you’re expected to do x”.

Dennett’s story outlines are, I suggest, the peculiarly human way to generalize about the world. In RCP, they amount to our cognitive style: a set of cognitive templates, which in its turn is the set of cognitive gadgets involved in exploiting the system’s own dynamics.

The importance that the templates comprising a cognitive style are limited in number should not be underestimated, for a huge number of cognitive templates would simply bring RP back.

This implies that the human-world comprises also templates rather than just kinds of situations.

Two quick remarks before going further on how cognitive style enables RP-free context switching: first, I see a cognitive style as a mechanism that is specialized for cultural evolution in Heyes’ sense (it provides means of enhancing the fidelity and accuracy with which information is passed on).

Second, I think that a cognitive style can be regarded as a metacognitive mechanism, at least in the conception advanceed by Joelle Proust.

Relevance-sensitive context switching

Here we’ll sketch two possible ways in which a system may switch from context to context: a bottom-up in which the system simply “finds itself” within a new context, and a top-down, in which the system can actively engage into bringing about a contextual tonality.

Perkins claims that behavior is largely controled by a bottom-up activity swiching mechanism he dubs *emergent activity switching*: it describes the dynamics of subsystems towards an attractor, just like incresing thirst eventually takes over and leads to water-seeking behavior.

The critical phase primes the system towards a certain contextual tonality. In RCP terms, this means it makes it enter a sequence of ec-patterns as described by a template.

This is a plausible account of how there can be stories that are always “trying to happen”. Different processes tend to build up to the point where some critical phase triggers a contextual tonality.

This pictures the system’s cognitive machinery as a dupe of such low-level undergoings, as if the background in which the system’s cognitive tasks take place was always a given, at least from the perspective of the system’s cognitive activity.

This brings forth a picture in which we don’t have to monitor deliberately whether it’s time to prepare for an interview (or a presentation, or a thesis defense). We can instead “trust our worry”. This is how our cognitive style interact with and preempt the cognitive tasks we put forth.

But there’s something unsatisfactory with this unidirectional account of the relation between contextual tonality and the cognitive tasks we engage in. Up to this point, this view pictures the system as a dupe of its own cognitive style. There are cases in which a system (like us) can actively engage in bringing the proper contextual tonality to the fore.

United Kingdom's George VI had not previous familiarity nor was prepared for being a king, which means almost nothing he was already familiar with could help him in the task of facing his stutter while giving a speech in the radio as king.

However, by training with preparatory routines, George could work out his own cognitive style and bring to the fore a new contextual tonality, and with it he brought up a distinct articulation of worldly knowledge and behavioral dispositions he could use to act differently in a very unfamiliar context.

The described process should not be taken as a full inner power that can fully overcome external circumstances or bring about a new contextual configuration out of the blue. Rather, we're pointing to a process of context-switching that is not a direct response to environmental changing, but a strategy actively adopted by the agent.

Closing remarks

Ultimately, relevance-sensitivity is the product of a cognitive gadget dubbed cognitive style, which is produced by cultural evolution and inherited through social learning.

Can a cognitive style handles non-saturable human-world contexts?

Yes, for we have, in a nutshell, a productive yet tamed capacity for context production that can extrapolate from the set of familiar contexts without implying that the involved cognitive mechanisms are able to spin freely in a void.

It is possible that we finally have what we need in order to have RP out of the way inasmuch as it bothers commonsense holism.

And the same goes for situation holism.

CONCLUSION

To conclude, let's recapitulate the big picture and then take stock. The two central theses presented in this dissertation were: (1) relevance sensitivity is a *cognitive gadget*; and (2) a representational account of this gadget can be provided by using frameworks compatible with *representational cognitive pluralism* (RCP).

Let's start with (1). I argued that we need culturally established cognitive styles in order to handle relevance in a culturally enriched environment. But by "enriched environment" I don't mean simply one that's embedded in any kind of culture. After all, in a sufficiently loose sense, many non-human animals can be said to inhabit culturally enriched environments. However, human culture is known to have peculiarities, such as its cumulative character. This enables humans to reach an unprecedented level of complexity and richness in their behavioral outputs. If indeed we're the only animal with cumulative culture, then it's likely that we're the only animal in need of culturally assembled cognitive styles. That's one of the reasons why simply accounting for cognition's flexibility is not enough. Showing that some system is able to exhibit a complex and flexible trajectory within the space of its possible states is necessary, but far from sufficient.

The case for the non-saturable character of human contexts is a way to shed some light on this. Humans can cope with an open-ended set of contexts. And each of these contexts has a non-saturable character, i.e. pretty much any feature within reach of the system's integrated capacities can impact the resulting behavior. This is why thinking about contexts as subsets of features is so misleading. Even absent features can play a role through their absence, and this cannot be modeled as the absence of a representation (or indication) for that feature in the model. That's also why we can't simply build a huge look-up table depicting every single possible context. And after a while, it should be somewhat evident that thinking about the human world in terms of types of situations leads to a dead end. At least while thinking about RP. It should be, but it's not (just remember Wheeler's insistence of his accounts of m-contexts through SPACs). That of course doesn't mean that there's no possible role for typified situations. It does mean that they can't tell the whole story, though.

The missing dimension in this story, as suggested, comprises the broad templates that characterize a cognitive style. They can "cut out" the world in terms of story templates or outlines. Such stories do not aim at describing specific situations, but the dynamics of broader aspects of human life (work, health, relationships, etc.) That's why they can be so relatively few and yet ground such a rich and open-ended set of behavioral outputs. They are always "trying to happen" in the background of the way we cope with the world, just like their counterparts within the cognitive machinery are (according to the presented suggestion), always "trying to happen". In this sense, a cognitive style can guide and tame the system. It does so by both helping it to figure out which cognitive assets must be employed at any given moment, and by helping it to exploit those assets with flexibility, but without going astray.

So, is RP solved? I wouldn't dare to say so. The issue affects a large portion of human cognition. There are just too many places from which something groundbreaking can emerge and require an overhaul of our way of thinking. As the cognitive sciences develop, one has to make sure that science and philosophy remain in tune. Philosophical reflections must be driven by the needs and traits of the relevant sciences, not the other way around. Thus, one must always be ready to rethink one's understanding of some phenomenon. Having said that, I do think there are reasons for being cautiously optimistic about making RP less mysterious. And I did my best to present these reasons as the discussion progressed. But of course, time and further empirical research will have to play their part. Therefore, in what follows, rather than making claims I can't fully ground, I'd rather talk about this work's contributions. There are some things that I regard as important steps towards a full-fledged account of relevance sensitivity. And at least with regard to these steps, I believe that this work makes (potentially) significant improvements.

First, this dissertation goes deeper and beyond most of the works dedicated to RP. For instance, relevance issues were sometimes regarded as problems grounded in the system's lack of flexibility. This is likely due to the rigid and static nature that mind mechanisms take in classic accounts of cognition (e.g. Fodor's idea that the mind comprises a limited number of well-defined specialized modules; minskyan frames, etc.). Thus, some think that by showing how a system can be flexible and enact a broad self-restructuring, one has already shown everything that's needed. As we saw in chapter 4, that's not at all the case. Showing that a system is capable of traversing many trajectories fluidly and flexibly is necessary, but not sufficient. Inasmuch as RP is concerned, what really matters is why the system took a certain trajectory rather than the other. That's what a cognitive style is supposed to tell us.

Furthermore, as exemplified in chapter 1, sometimes the problem is taken as that of being able to navigate what one already knows, i.e. what one is already familiar with. But this amounts to a rather inaccurate characterization of the phenomenon. Grasping what's relevant cannot simply rely on the knowledge obtained from encounters with previously familiar situations. That is problematic not just because it raises the bootstrap question: what about the first time in which the agent faced that situation? It is problematic because, as exhaustively discussed, even situations of the same kind can vary drastically. Avoiding this mistake is what motivated the depiction of commonsense as the ability to extrapolate from what one already knows. A similar point can be found in Dreyfus; Dreyfus (1987). They were discussing the jockey example presented in chapter 1: a horse racing gambler who's trying to figure out her bet learns that a given jockey had hay-fever and that the race-course landscaping is in full-flower. How can we accommodate the fact that this emerges as immediately relevant for us, even though facts about health are usually decoupled from facts about horse-race betting? Here is what they say about it:

> Data structures of preclassified, commonsense facts dear to AI researchers cannot account for this amazing capacity to see *new* relevance; but our claim that through experience each person has embodied a great deal of commonsense know-how does not help either. Such new facts would not be included in a frame for horse-race betting, but neither would they show up as salient

> for a skilled human being, even though his past went ahead of him organizing his world. There is a sort of everyday creativity which seems to be equally beyond the grasp of cognitivism and of phenomenology. (Dreyfus; Dreyfus, 1987, p. 111)

Quite curiously though, the same Dreyfus seems to have forgotten his own insight. In later work, he praises dynamic models as a viable alternative to computational ones. And he does that for thinking that these models exhibit the necessary flexibility to reorganize the world according to the system's previous experiences (Dreyfus, 2007). As for me, I think the need to handle what Dreyfus dubbed "new relevance" is one of the most important insights regarding RP. It is an essential feature of our sensitivity to relevance. And yet it seems like a bunch of important work (including that of Dreyfus himself, Wheeler, Vervaeke and others) don't give it enough importance. Emphasizing this trait and providing an explicit account for it is, I believe, another important contribution of this work. Cognitive styles are not just routines or situations with which the system is already familiar: it comprises also a way to exploit their permutations while still constrained by relevance sensitivity.

Second, the work inherits some purchases of Heyes' cognitive gadgets framework. It brings into focus (and renders tractable) questions about how cognitive mechanisms are assembled over time. Though some cognitive tools can be understood in isolation of their origin and formation, that does not seem to be the case with relevance sensitivity. Indeed, an important lesson of the discussion we put forth is that if we don't have a story to tell about the mechanism's origin, it might remain functionally mysterious. This is the kind of emphasis that we find in the cognitive gadgets framework:

> Evolutionary psychologists tend to assume that, if something is a cognitive instinct, it is the responsibility of some other discipline — perhaps genetics or paleoarchaeology, but not cognitive science — to explain how it was constructed. In contrast, cultural evolutionary psychology encourages cognitive scientists and others to develop and test theories about how cognitive gadgets are put together over time. (Heyes, 2018a, p. 222)

The upshot is that Heyes' framework can not only buy us a way out of a hard issue. It also helps us to keep the problem always on the horizon of our empirical research. Relevance sensitivity cannot remain in the background.

What about (2)? RCP is an alternative answer to the challenge of representational productivity. It enables a broadly representational account of cognition that won't suffer from FP. But it does so not by taking for granted that cognition is broadly representational. Rather, it allows the formulation of empirical hypotheses involving representations. The point is that, even if it turns out that cognition is broadly representational, the large-scale articulation of representational assets won't bring FP back with it. RCP can buy us that because it employs a plurality of domain-targeted structural schemes. And the expressive power of these domain-specific schemes can be gradually enhanced through representational redescription.

Within this work, the main goal of RCP is to enable the empirical possibility of a broadly representational account of cognitive styles. Non-representational frameworks need not worry about FP, but in order to do that, they must give up a potentially valuable cognitive asset (i.e. representations). RCP can overcome FP, which is unsolvable in representational systems that rely on the absolute contents of language-like schemes. And it does so without giving up on representational productivity. Thus, what seemed to be a disadvantage of representational frameworks (i.e. FP) towards handling RP becomes a strength. We can have the explanatory purchases that representational contents afford without worrying about foundational issues like FP.

Now, it is true that those who don't like the story as I told may have an easy time picking up just the points they like (if any) and commit the rest to the flames. One can, for instance, tell a story of what a cognitive style is or does within another framework. Perhaps a friend of massive modularity might think that cognitive styles can be stored in something like Sperber's physiological markers. Or maybe a friend of Wheeler's heideggerian approach regards styles as drawing on the activation patterns of the many CRC mechanisms comprising the system. Providing this kind of account is certainly a plausible goal. Throughout the research, I considered these possibilities myself. What took me away from them is that they would bring to the table predicaments and difficulties that do not emerge within RCP, specially when considering offline capacities. After all, the knowledge stored as physiological markers and environmental couplings must be somehow rendered portable and available for further exploitation (consider again the examples of top-down context switching). A representational approach strikes me as the one with the best trade-off between theoretical commitments and explanatory purchases. That makes it worth pursuing as much as allowed by the available evidence. Thus, my point is not that there's a knock-down argument against these non-representational alternatives, only that I think they would result in a less flexible and less powerful framework (assuming they can be made to work, of course).

Despite its role in accounting for cognitive styles, RCP has its own merits, and its potential contributions can go beyond that. First, it is broadly compatible with contemporary work within cognitive sciences (including cognitive neurosciences). For instance, one need not give up on any core tenet of situated cognition. Indeed, I'm an advocate of situated cognition myself, and anything incompatible with it would certainly bother me. But even those still wary of situated cognition — for whatever reason I don't imagine — can still benefit from RCP. That is so because, as discussed, the main differences among frameworks concern the dimension of target fixation, not representational contents. Thus, if anyone wants to use RCP within a broadly non-situated approach, so be it. She'll have to deal with the empirical difficulties stemming from her outmoded approach, but RCP will not be in her way.

As I see it, this framework-wise flexibility is a very important trait. An approach that enables the formulation of empirical hypotheses involving certain cognitive assets (representations, m-contexts, etc.) should be preferred over one that promptly assumes or rejects them *a priori*. This is why, for instance, I do not try to provide a direct argument for the necessity of representations. I'm not interested in reenacting Andy Clark's argument for "representation-hungry" cognitive tasks (1997a). My goal is to allow both friends and foes of representational contents to formulate empirical hypotheses regarding their existence and role in explanations of specific cognitive capacities or performances. That must be done for each and every capacity we're interested in explaining. And in each case, may the best available explanation wins.

Another significant contribution of RCP comes from the extra degrees of freedom in the formulation of empirical hypotheses. The first is a gift from Cummins' theory of representational contents. It provides a purely semantic account of misrepresentation, enabling us to conceive of well-functioning mechanisms under ideal environmental circumstances that nonetheless misinform. In other words, we have an account of misrepresentation that need not boil down to malfunction, inappropriate circumstances, inapt behavior and the like. This makes Cummins' theory of representational contents quite resilient versus the most well-known arguments against representations: it overcomes classic naturalization challenges, and they can play an evidently non-trivial explanatory role.

The second extra dimension unfolded by RCP comes from its reliance on structural representations. They can represent the dynamics of some domain, rather than just computing it. In other words, rather than representing the current state and compute its changes, the change itself can be "pictured". This is quite a feature, and it plays a fundamental role in the account of cognitive gadgets provided here. After all, templates for dynamics are nothing but representations of that dynamics. A representational account of cognitive styles would not be possible without it, for any language-like account of such dynamics would inevitably suffer from FP. Needless to say, this feature of structural representations is available for explanations of other cognitive capacities as well.

The third remarkable extra degree of freedom is in the huge flexibility with which we can model distinct cognitive systems. LOT-powered systems imply that anything endowed with cognitive capacities must be assembled with a specific architecture. In other words, all capacity potentially cognitive must be in a single continuum towards increasingly powerful versions of that same architecture. That makes it hard to account for discrepancies such as those we find between human and non-human animals. This quite inflexible picture of the cognitive machinery is inevitably connected with the tendency to be excessively reluctant or excessively inclined to attribute cognitive capacities to non-human animals. Consider, for instance, how Povinelli is sometimes misrepresented as someone who denies cognitive capacities for apes.[24] All he does is deny that apes' cognition is human-like. But of course if there's no clear choice between the Scylla of behaviorism and the Charybdis of a LOT architecture, then whenever we see apes behaving like we do, we'll be certainly tempted to attribute them cognitive mechanisms like ours.

My point here is just this: RCP sheds light on a dimension encompassing very distinct cognitive capacities whose underlying machinery can nonetheless be put on the same evolutionary continuum. Examples abound: ape cognition and human cognition may be distinct in

[24] He talks a bit about it in Povinelli (2012).

how much they rely on representations rather than, e.g. encodings or indicator signals. They might also imply distinct structural schemes, or perhaps exploit them differently. Another possibility is that they may differ in how much (or if) they can engage in representational redescription. Thus, even if humans share some architectural feature with apes (as is likely the case), humans might be the ones that, say, learned how to task-embed stuff from other domains within it. All of these possibilities are available for investigation within a framework compatible with RCP. That's what make it a powerful choice to account for the differences and similarities within the cognitive apparatuses of any two species.

There's one last contribution I'd like to remark. It concerns neither (1) nor (2) in isolation, but rather their conjunction. That is, it emerged as the result of pursuing both at the same time. It's about the role of context in cognitive explanations, i.e. the role of context *qua* explanatory resource rather than a capacity to be explained. This is what I called *m-context*. In most of the relevant literature, the potential explanatory role of m-contexts was conflated with the potential explanatory role of representational contents. As discussed in the first chapter, this conflation is widespread since Dreyfus' early critique of AI. In his view, m-contexts are useless because they necessarily amount to some kind of data structure used to represent a context (e.g. a minskyan frame). But this amounts to thinking the role of m-contexts within the boundaries of a very specific framework. It is true that, for Dreyfus, modeling contexts as data structures was the only game in town. But even when new games started to pop out, the conflation stayed on. Even today, it is not hard to find researchers redirecting the whole discussion from one about m-contexts to one about the expressive power of the employed representational scheme. Such a proposal clearly conflates issues about representational productivity with issues of inferential productivity. But if the discussion in the first chapter contains any lesson, it is that the potential roles and caveats of m-contexts happens also in broadly non-representational approaches like Wheeler's. This allowed me to avoid taking a discussion about the potential explanatory role of m-contexts to be a discussion about whether m-contexts can be representations of contexts.

Given the distinction, one can finally concentrate on the potential explanatory role of m-contexts. In standard stories, the role of m-contexts is that of being a mechanistic counterpart for the system's current context. That is, we can explain the system's capacity to properly behave within a given context (say restaurant) in terms of some internal organization of the relevant cognitive knowledge involved in handling restaurants (i.e. a restaurant m-context). But as we've seen, cognitive styles need not rely on m-contexts in order to play their role. Cognitive styles are not collections of structured m-contexts, but rather collections of large-scale templates of the cognition's dynamics, and collections of strategies employed when exploiting those templates. The upshot is that the potential reliance on m-contexts, once considered the heart of the matter, was revealed as tangential to RP. Thus, the possibility arises that, just like the explanatory role of representations, the role of m-contexts (if any) is also a matter for empirical research. For instance, consider again the discussion between Rietveld and Wheeler around patients of utilization-behavior that took place in chapter 1. Their disagreement was essentially about whether m-context has a role to play in the explanation of that condition. Whatever turns out to be the best account, it can be easily accommodated within RCP, and it won't be a problem for (1).

This means that neither (1) nor (2) implies a commitment to neither the reliance nor the rejection of m-contexts. In this sense, both theses are attempts to rely on cognitive assets quite distinct from those traditionally employed by most researchers. As a consequence, even if they turn out to be wrong, it's going to be for a new reason. In my view, that's what makes philosophical endeavors so interesting: to find a new way to be wrong may also bear fruits.

BIBLIOGRAPHY

ABNAR, S. et al. Blackbox meets blackbox: Representational similarity and stability analysis of neural language models and brains. 2019.

ANDERSEN, B. P.; MILLER, M.; VERVAEKE, J. Predictive processing and relevance realization: Exploring convergent solutions to the frame problem. **Phenomenology and the Cognitive Sciences**, 2022.

ANDERSON, M. L. Neural reuse: A fundamental organizational principle of the brain. **Behavioral and Brain Sciences**, v. 33, n. 4, p. 245–266, 2010.

ANDERSON, M. L. **After phrenology: Neural reuse and the interactive brain**. [s.l.] MIT Press, 2014.

ARJOVSKY, M. Out of Distribution Generalization in Machine Learning. **arXiv:2103.02667 [cs, stat]**, 2021.

BARLOW, H. B. Possible principles underlying the transformations of sensory messages. In: ROSENBLITH, W. A. (Ed.). **Sensory communication**. [s.l: s.n.]. p. 217–234.

BARLOW, H. B. Unsupervised learning. **Neural Computation**, v. 1, n. 3, p. 295–311, 1989.

BARRETT, L. Out of their heads: Turning relational reinterpretation inside out. **Behavioral and Brain Sciences**, v. 31, n. 2, p. 130–131, 2008.

BARSALOU, L. W. Perceptions of perceptual symbols. **Behavioral and Brain Sciences**, v. 22, n. 4, p. 637–660, 1999.

BARSALOU, L. W. Abstraction in perceptual symbol systems. **Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences**, v. 358, n. 1435, p. 1177–1187, 2003.

BARTH, C. H. **O frame problem: A sensibilidade ao contexto como um desafio para teorias representacionais da mente**. Master's thesis—Belo Horizonte: Faculdade de Filosofia e Ciências Humanas, Universidade Federal de Minas Gerais, 2018.

BARTH, C. H. É possível evitar vieses algorítmicos. **Revista de Filosofia Moderna e Contemporânea**, v. 8, n. 3, p. 39–68, 2021.

BICKHARD, M. H. Why children don't have to solve the frame problems: Cognitive representations are not encodings. **Developmental Review**, v. 21, n. 2, p. 224–262, 2001.

BLACKMON, J. et al. Systematicity and the cognition of structured domains. **The Journal of Philosophy**, v. 98, n. 4, p. 167, 2001.

BOERLIN, M.; DENÈVE, S. Spike-based population coding and working memory. **PLoS Computational Biology**, v. 7, n. 2, p. e1001080, 2011.

BROOKS, R. A. Intelligence without representation. **Artificial Intelligence**, n. 47, p. 139–159, 1991.

BRUINEBERG, J.; CHEMERO, A.; RIETVELD, E. General ecological information supports engagement with affordances for "higher" cognition. **Synthese**, v. 196, n. 12, p. 5231–5251, 2018.

BRUINEBERG, J.; RIETVELD, E. Self-organization, free energy minimization, and optimal grip on a field of affordances. **Frontiers in Human Neuroscience**, v. 8, 2014.

BURGESS, N.; O'KEEFE, J.; RECCE, M. Using hippocampal 'place cells' for navigation, exploiting, phase coding. In: HANSON, S. J.; COWAN, J. D.; MILES, C. L. (Eds.). **Neural information processing systems 5**. San Mateo: Morgan Kaufmann, 1993. p. 929–936.

CALDIEU, C.; OLSHAUSEN, B. Learning transformational invariants from natural movies. **Proceedings of the 21st International Conference on Neural Information Processing Systems**, p. 209–216, 2008.

CAMP, E. Why maps are not propositional. In: GRZANKOWSKI, A.; MONTAGU, M. (Eds.). **Non-propositional intentionality**. Oxford: Oxford, 2018. p. 20–45.

CAMPBELL, J. **The hero with a thousand faces**. [s.l.] Pantheon Books, 1949.

CAO, R. New labels for old ideas: Predictive processing and the interpretation of neural signals. **Review of Philosophy and Psychology**, v. 11, n. 3, p. 517–546, 2020.

CAPPUCCIO, M.; WHEELER, M. Ground-level intelligence: Action-oriented representation and the dynamics of the background. In: **Knowing without thinking**. [s.l.] Palgrave Macmillan UK, 2012. p. 13–36.

CARRUTHERS, P. On Fodor's Problem. **Mind and Language**, v. 18, n. 5, p. 502–523, 2003.

CARVALHO, E. M. DE; ROLLA, G. O desafio da integração explanatória para o enativismo: Escalonamento ascendente ou descendente. **Prometheus**, n. 33, 2020a.

CARVALHO, E. M. DE; ROLLA, G. An enactive-ecological approach to information and uncertainty. **Frontiers in Psychology**, v. 11, 2020b.

CARVALHO, F. N. DE. O papel do contexto na percepção das emoções. **Perspectiva Filosófica**, v. 46, n. 2, p. 116–142, 2019.

CARVALHO, F. N. DE. Fearful object seeing. **Review of Philosophy and Psychology**, 2021.

CATMUR, C.; WALSH, V.; HEYES, C. Associative sequence learning: The role of experience in the development of imitation and the mirror system. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 364, n. 1528, p. 2369–2380, 2009.

CHEMERO, A. **Radical embodied cognitive science**. Cambridge, MA: MIT Press, 2009.

CHEMERO, A.; SILBERSTEIN, M. After the philosophy of mind: Replacing scholasticism with science. **Philosophy of Science**, v. 75, n. 1, p. 1–27, 2008.

CHERNIAK, C. **Minimal rationality**. [s.l.] MIT Press, 1990.

CHIAPPE, D. L.; KUKLA, A. Context selection and the frame problem. **Behavioral and brain sciences**, v. 19, n. 3, p. 529–530, 1996.

CHOW, S. J. What's the problem with the frame problem? **Review of Philosophy and Psychology**, v. 4, n. 2, p. 309–331, 2013.

CHURCHLAND, P. M. **A neurocomputational perspective: The nature of mind and the structure of science**. Cambridge: The MIT Press, 1989.

CHURCHLAND, P. M. Conceptual similarity across sensory and neural diversity: The fodor/lepore challenge answered. **The Journal of Philosophy**, v. 95, n. 1, p. 5, 1998.

CLARK, A. **Associative engines: Connectionism, concepts, and representational change**. Cambridge: The MIT Press, 1993.

CLARK, A. The dynamical challenge. **Cognitive Science**, v. 21, n. 4, p. 461–481, 1997b.

CLARK, A. **Being there: Putting brain, body, and world together again**. Cambridge: MIT Press, 1997a.

CLARK, A. Local Associations and Global Reason: Fodor's Frame Problem and Second-Order Search. 2002.

CLARK, A. **Surfing uncertainty**. Oxford: Oxford University Press, 2016.

CLARK, A.; THORNTON, C. Trading spaces: Computation, representation, and the limits of uninformed learning. **Behavioral and Brain Sciences**, v. 20, p. 57–90, 1997.

CRICK, F.; KOCH, C. A framework for consciousness. **Nature Neuroscience**, n. 6, p. 119–126, 2003.

CRYSTAL, J. D. Remembering the past and planning for the future in rats. **Behavioural Processes**, v. 93, p. 39–49, 2013.

CUMMINS, R. **The nature of psyhocological explanation**. Cambridge: The MIT Press, 1983.

CUMMINS, R. Conceptual Role Semantics and the Explanatory Role of Content. **Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition**, v. 65, n. 1/2, p. 103–127, 1992.

CUMMINS, R. **Representations, targets and attitudes**. [s.l.] MIT Press, 1996.

CUMMINS, R. Reply to millikan. **Philosophy and Phenomenological Research**, v. 60, n. 1, p. 113, 2000.

CUMMINS, R. The LOT of the causal theory of mental content. In: **The world in the head**. [s.l.] Oxford, 2010c. p. 11–19.

CUMMINS, R. Representational specialization: The synthetic a priori revisited. In: CUMMINS, R. (Ed.). **The world in the head**. [s.l.] Oxford, 2010b. p. 194–209.

CUMMINS, R. et al. Representation and unexploited content. In: **The world in the head**. [s.l.] Oxford, 2010. p. 120–133.

CUMMINS, R. How does it work versus ”what are the laws?”: Two conceptions of psychological explanation. In: CUMMINS, R. (Ed.). **The world in the head**. [s.l.] Oxford, 2010a. p. 283–310.

CUMMINS, R. C. Cross-domain inference and problem embedding. In: CUMMINS, R. E.; POLLOCK, J. L. (Eds.). **Philosophy and AI: Essays at the interface**. [s.l.] MIT Press, 1991.

CUMMINS, R.; POIRIER, P. Representation and indication. In: **The world in the head**. [s.l.] Oxford, 2010. p. 98–119.

CUMMINS, R.; POIRIER, P.; ROTH, M. Epistemological strata and the rules of right reason. **Synthese**, v. 141, n. 3, p. 287–331, 2004.

DAMASIO, A.; DAMASIO, H. Cortial systems for retrieval of concrete knowledge: The convergence zone framework. In: KOCH, C.; DAVIS, J. (Eds.). **Large-Scale Neuronal Theories of the Brain**. [s.l.] MIT Press, 1994.

DE WAAL, F. **Are we smart enough to know how smart animals are?** [s.l.] W. W. Norton & Co., 2016.

DEHAENE, S. **How we learn**. [s.l.] Penguin Books, 2021.

DENNETT, D. Inteligência artificial como filosofia e como psicologia. In: **Brainstorms: Ensaios filosóficos sobre a mente e a psicologia**. Translation: Luiz Henrique De Araújo Dutra. [s.l.] Unesp, 1999[1978]. p. 163–183.

DENNETT, D. Artificial intelligence as philosophy and as psychology. In: **Brainstorms: Philosophical essays on mind and psychology**. [s.l.] MIT Press, 1981.

DENNETT, D. Cognitive wheels: The frame problem of AI. In: PYLYSHYN, Z. W. (Ed.). **The robot’s dillemma: The frame problem in artificial intelligence**. [s.l.] Ablex, 1987. p. 41–64.

DENNETT, D. Mother nature vs. The walking encyclopedia. In: RAMSEY, W.; STITCH, S. P.; RUMELHART, D. E. (Eds.). **Philosophy and Connectionist Theory**. New Jersey: Lawrence Erlbaum Associates, 1991.

DENNETT, D. Producing future by telling stories. In: FORD, K. M.; PYLYSHYN, Z. W. (Eds.). **The robot’s dillemma revisited: The frame problem in artificial intelligence**. [s.l.] Ablex, 1996. p. 1–8.

DOMINGOS, P. **The master algorithm : How the quest for the ultimate learning machine will remake our world**. New York: Basic Books, 2015.

DRETSKE, F. **Knowledge and the flow of information**. Cambridge: MIT Press, 1981.

DRETSKE, F. **Explaining behavior: Reasons in a world of causes**. [s.l.] The MIT Press, 1991.

DRETSKE, F. I. Misrepresentation. In: BOGDAN, R. (Ed.). **Belief: Form, content, and function**. [s.l.] Oxford University Press, 1986. p. 17–36.

DREYFUS, H. **What computers can't do: A critique of artificial reason**. New York: Harper & Row, 1972.

DREYFUS, H. **What computers still can't do**. [s.l.] MIT Press, 1992.

DREYFUS, H. L. Why heideggerian AI failed and how fixing it would require making it more heideggerian. In: **Artificial Intelligence**. [s.l.] Elsevier, 2007. v. 171, p. 1137–1160.

DREYFUS, H. L.; DREYFUS, S. E. How to stop worrying about the frame problem even though it's computationally insoluble. In: PYLYSHYN, Z. W. (Ed.). **The robot's dillemma: The frame problem in artificial intelligence**. [s.l.] Ablex, 1987. p. 95–111.

EGAN, F. A deflationary account of mental representation. In: SMORTCHKOVA, J.; DOLEGA, K.; SCHLICHT, T. (Eds.). **Mental representations**. [s.l.] New York, USA: Oxford University Press, 2020.

ESSEN, D. C. V.; ANDERSON, C. H.; OLSHAUSEN, B. A. Dynamic Routing Strategies in Sensory, Motor, and Cognitive Processing. In: KOCH, C.; DAVIS, J. (Eds.). **Large-Scale Neuronal Theories of the Brain**. [s.l.] MIT Press, 1994.

FACCHIN, M. Structural representations do not meet the job description challenge. **Synthese**, v. 199, n. 3-4, p. 5479–5508, 2021.

FARIES, F.; CHEMERO, A. Dynamic information processing. In: **The routledge handbook of the computational mind**. [s.l.] Taylor
& Francis Ltd, 2018. p. 134–148.

FODOR, J. Having concepts: A brief refutation of the twentieth century. **Mind and Language**, v. 19, n. 1, p. 29–47, 2004.

FODOR, J. A. **The language of thought**. [s.l.] Harvard University Press, 1980.

FODOR, J. A. **Representations**. Sussex: The Harvester Press, 1981. v. 13.

FODOR, J. A. **The modularity of mind**. [s.l.] MIT Press, 1983.

FODOR, J. A. Modules, frames, fridgeons, sleeping dogs and the music of the spheres. In: PYLYSHYN, Z. W. (Ed.). **The robot's dillemma: The frame problem in artificial intelligence**. [s.l.] Ablex, 1987. p. 139–149.

FODOR, J. A. **Concepts: Where cognitive science went wrong (oxford cognitive science series)**. [s.l.] Oxford University Press, 1998.

FODOR, J. A. **LOT 2: The language of thought revisited**. [s.l.] Oxford University Press, 2010.

FODOR, J. A.; PYLYSHYN, Z. W. Connectionism and cognitive architecture: A critical analysis. **Cognition**, v. 28, n. 1-2, p. 3–71, 1988.

FODOR, J.; LEPORE, E. **Holism : A shopper's guide**. Oxford Cambridge, Mass., USA: Blackwell Publishers, 1991.

FODOR, J.; LEPORE, E. Paul churchland and state space semantics. In: MCCAULEY, R. N. (Ed.). **The churchlands and their critics**. Cambridge: Blackwell, 1996.

FORD, K. M.; PYLYSHYN, Z. W. (EDS.). **The robot's dilemma revisited: The frame problem in artificial intelligence**. Norwood, NJ, USA: Ablex Publishing Corp., 1996.

FREEMAN, W. J. **How brains make up their minds**. 1. ed. [s.l.] Columbia University Press, 2001.

FRISTON, K. Active inference and free energy. **Behavioral and Brain Sciences**, v. 36, n. 3, p. 212–213, 2013.

FRISTON, K. J. Functional and effective connectivity in neuroimaging: A synthesis. **Human Brain Mapping**, v. 2, n. 1-2, p. 56–78, 1994.

FRISTON, K.; MATTOUT, J.; KILNER, J. Action understanding and active inference. **Biological Cybernetics**, v. 104, n. 1-2, p. 137–160, 2011.

GALILEI, G. **Dialogues concerning two new sciences**. Translation: Henry Crewand;Translation: Alfonso De Salvio. [s.l.] Dover, 1954.

GALLAGHER, S. Review: Reconstructing the cognitive world: The next step. **Mind**, v. 116, n. 463, p. 792–796, 2007.

GEIRHOS, R. et al. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. **CoRR**, v. abs/1811.12231, 2019.

GIBSON, J. J. **The ecological approach to visual perception**. [s.l.] Houghton Mifflin, 1979.

GLYMOUR, C. Android epistemology and the frame problem: Comments on dennett's Cognitive Wheels. In: PYLYSHYN, Z. W. (Ed.). **The robot's dillemma: The frame problem in artificial intelligence**. [s.l.] Ablex, 1987. p. 65–75.

GREGOR, K.; LECUN, Y. Learning fast approximations of sparse coding. **Proceedings of the 27th International Confer-ence on Machine Learning**, p. 399–406, 2010.

GRIFFA, A. et al. The evolution of information transmission in mammalian brain networks. 2022.

HASELAGER, W. F. G. **Cognitive science and folk psychology**. London: Sage, 1997.

HATEREN, J. H. VAN; RUDERMAN, D. L. Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. **Proceedings of the Royal Society of London. Series B: Biological Sciences**, v. 265, n. 1412, p. 2315–2320, 1998.

HAUGELAND, J. **Artificial intelligence: The very idea**. [s.l.] Bradford, 1985.

HAUGELAND, J. An overview of the frame problem. In: PYLYSHYN, Z. W. (Ed.). **The robot's dillemma: The frame problem in artificial intelligence**. [s.l.] Ablex, 1987. p. 76–93.

HAUGELAND, J. Body and world: A review of what computers still can't do: A critique of artifcial reason. **Artificial intelligence**, v. 80, n. 1, p. 119–128, 1996.

HAUGELAND, J. The intentionality all-starts. In: **Having thought**. Cambridge: Harvard University Press, 1998c. p. 128–170.

HAUGELAND, J. Truth and rule-following. In: **Having thought**. Cambridge: Harvard University Press, 1998f. p. 305–361.

HAUGELAND, J. Understanding natural language. In: **Having thought**. Cambridge: Harvard University Press, 1998a. p. 47–60.

HAUGELAND, J. The nature and plausibility of cognitivism. In: HAUGELAND, J. (Ed.). **Having thought**. Cambridge: Harvard University Press, 1998b. p. 9–45.

HAUGELAND, J. Representational genera. In: **Having thought**. Cambridge: Harvard University Press, 1998e. p. 171–206.

HAUGELAND, J. Mind embodied and embedded. In: HAUGELAND, J. (Ed.). **Having thought**. Cambridge: Harvard University Press, 1998d. p. 207–237.

HAYES, P. J. **In defence of logic**. Proceedings IJCAI 77. p. 559–5651977.

HAYES, P. J. What the frame problem is and isn't. In: PYLYSHYN, Z. W. (Ed.). **The robot's dillemma: The frame problem in artificial intelligence**. [s.l.] Ablex, 1987. p. 123–137.

HEIDEGGER, M. **Ser e tempo**. Translation: Fausto Castilho. Campinas: Unicamp, 2012.

HENDRICKS, S. The frame problem and theories of belief. **Philosophical studies**, v. 129, n. 2, p. 317–333, 2006.

HEYES, C. Grist and mills: On the cultural origins of cultural learning. **Philosophical transactions of the Royal Society**, v. 367, n. 1599, p. 2181–2191, 2012.

HEYES, C. **Cognitive gadgets**. [s.l.] Harvard University Press, 2018a.

HEYES, C. Empathy is not in our genes. **Neuroscience and Biobehavioral Reviews**, v. 95, p.

499–507, 2018b.

HEYES, C. et al. Knowing ourselves together: The cultural origins of metacognition. **Trends in Cognitive Sciences**, v. 24, n. 5, p. 349–362, 2020.

HOHWY, J. **The predictive mind**. [s.l.] Oxford University Press, 2014.

HRDY, S. **Mothers and others**. [s.l.] Harvard University Press, 2011.

HUBEL, D. H.; WIESEL, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. **Journal of phisiology**, n. 160, p. 106–154, 1962.

HUME, D. **A treatise of human nature**. [s.l.] Oxford University Press, 1965.

HURLEY, S. Perception and action: Alternative views. **Synthese**, v. 129, n. 1, p. 3–40, 2001.

HURLEY, S. Making sense of animals. In: HURLEY, S.; NUDDS, M. (Eds.). **Rational animals?** [s.l.] Oxford, 2006. p. 139–171.

HUTTER, M.; LEGG, S. A collection of definitions of intelligence. 2007.

HUTTO, D.; MYIN, E. **Radicalizing enactivism : Basic minds without content**. Cambridge, Mass: MIT Press, 2013.

HUYS, Q. J. M. et al. Bonsai trees in your head: How the pavlovian system sculpts goal-directed choices by pruning decision trees. **PLoS Computational Biology**, v. 8, n. 3, p. e1002410, 2012.

HUYS, Q. J. M. et al. Interplay of approximate planning strategies. **Proceedings of the National Academy of Sciences**, v. 112, n. 10, p. 3098–3103, 2015.

ISMAEL, J. T. The situated self. 2007.

JACK, R. E. et al. Cultural confusions show that facial expressions are not universal. **Current Biology**, v. 19, n. 18, p. 1543–1548, 2009.

JANLERT, L.-E. The frame problem: Freedom or stability? With pictures we can have both. In: FORD, K. M.; PYLYSHYN, Z. W. (Eds.). **The robot's dillemma revisited: The frame problem in artificial intelligence**. [s.l.] Ablex, 1996. p. 35–48.

KAHNEMAN, D. **Thinking, Fast and Slow**. [s.l.] Farrar, Straus; Giroux, 2011.

KAPLAN, C. A.; SIMON, H. A. In search of insight. **Cognitive Psychology**, v. 22, n. 3, p. 374–419, 1990.

KARKLIN, Y.; LEWICKI, M. S. Emergence of complex cell properties by learning to generalize in natural scenes. **Nature**, v. 457, n. 7225, p. 83–86, 2008.

KARMILOFF-SMITH, A. **Beyond modularity : A developmental perspective on cognitive science**. Cambridge, Mass: MIT Press, 1992.

KIVERSTEIN, J. D.; RIETVELD, E. Reconceiving representation-hungry cognition: An ecological-enactive proposal. **Adaptive Behavior**, v. 26, n. 4, p. 147–163, 2018.

KIVERSTEIN, J.; WHEELER, M. **Heidegger and cognitive science**. New York: Palgrave Macmillan, 2012.

KLEINBERG, J. M. Authoritative sources in a hyperlinked environment. **Journal of the ACM**, v. 46, n. 5, p. 604–632, 1999.

LAAKSO, A.; COTTRELL, G. Content and cluster analysis: Assessingrepresentational similarity in neuralsystems. **Philosophical Psychology**, v. 13, n. 1, p. 47–76, 2000.

LEWICKI, M. S.; OLSHAUSEN, B. A. Probabilistic framework for the adaptation and comparison of image codes. **Journal of the Optical Society of America A**, v. 16, n. 7, p. 1587, 1999.

LORMAND, E. The holorobophobe's dilemma. In: FORD, K. M.; PYLYSHYN, Z. W. (Eds.). **The robot's dillemma revisited: The frame problem in artificial intelligence**. [s.l.] Ablex, 1996. p. 61–88.

MACHERY, E. **Doing without concepts**. Oxford New York: Oxford University Press, 2009.

MARCUS, G. **GPT-2 and the Nature of Intelligence**. **The Gradient**, 2020. Available at: https://thegradient.pub/gpt2-and-the-nature-of-intelligence/. Access date: 26 may. 2020

MARCUS, G. F. Rethinking eliminative connectionism. **Cognitive Psychology**, v. 37, 1998.

MARCUS, G. F. **The algebraic mind: Integrating connectionism and cognitive science**. Cambridge: MIT press, 2003.

MARCUS, G.; DAVIS, E. **Rebooting AI**. New York: Pantheon, 2019.

MARCUS, G.; DAVIS, E. **GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about**. **MIT Technology Review**, 2020. Available at: https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/. Access date: 25 sep. 2020

MARR, D. **Vision: A computation investigation into the human representational system and processing of visual information**. San Francisco, CA: MIT Press Ltd, 2010.

MCCARTHY, J. Circumscription - a form of non-monotonic reasoning. **Artificial intelligence**, v. 13, n. 1, p. 27–39, 1980.

MCCARTHY, J.; HAYES, P. J. Some philosophical problems from the standpoint of artificial intelligence. **Machine Intelligence**, v. 4, p. 463–502, 1969.

MCCLELLAND, J. L. et al. **Parallel distributed processing, vol. 2: Psychological and biological models**. Cambridge, MA: MIT press, 1987. v. 2.

MCCOY, R. T. et al. Embers of autoregression: Understanding large language models through the problem they are trained to solve. 2023.

MCDERMOTT, D. Artificial intelligence meets natural stupidity. **ACM SIGART Bulletin**, n. 57, p. 4–9, 1976.

MCDERMOTT, D. We've been framed: Or, why AI is innocent of the frame problem. In: PYLYSHYN, Z. W. (Ed.). **The robot's dillemma: The frame problem in artificial intelligence**. [s.l.] Ablex, 1987. p. 113–122.

MCDERMOTT, D.; DOYLE, J. Non-monotonic logic i. **Artificial intelligence**, v. 13, n. 1-2, p. 41–72, 1980.

MCDOWELL, J. The content of perceptual experience. **The Philosophical Quarterly**, v. 44, n. 175, p. 190, 1994.

MCGURK, H.; MACDONALD, J. Hearing lips and seeing voices. **Nature**, v. 264, n. 5588, p. 746–748, 1976.

MERLEAU-PONTY, M. **Fenomenologia da percepção**. Translation: Carlos Alberto Ribeiro De Moura. São Paulo: Martins Fontes, 2011.

MILLIKAN, R. G. **Language, thought, and other biological categories: New foundations for realism**. [s.l.] The MIT Press, 1987.

MINSKY, M. A framework for representing knowledge. In: HAUGELAND, J. (Ed.). **Mind design II: Phylosophy, psychology, artificial intelligence**. [s.l.] MIT Press, 1997. p. 111–142.

NAGEL, T. **The view from nowhere**. New York: Oxford University Press, 1986.

NANDA, V. et al. Measuring representational robustness of neural networks through shared invariances. 2022.

NEWELL, A.; SHAW, J. C.; SIMON, H. A. **Report on a general problem solving program**. IFIP congress. v. 256, p. 64Pittsburgh, PA, 1959.

NEWELL, H. A., Allen; Simon. Computer science as empirical inquiry: Symbols and search. **Communications of the ACM**, v. 19, 1976.

NEWEN, A.; VOSGERAU, G. Situated mental representations: Why we need mental representations and how we should understand them. In: SMORTCHKOVA, J.; DOLEGA, K.; SCHLICHT, T. (Eds.). **What are mental representations?** [s.l.] Oxford University Press, 2020. p. 178–212.

NOË, A. **Action in perception**. Cambridge: MIT Press, 2004.

NOË, A. **Varieties of presence**. Cambridge: Harvard University Press, 2012.

OKEEFE, J.; BURGESS, N. Dual phase and rate coding in hippocampal place cells: Theoretical significance and relationship to entorhinal grid cells. **Hippocampus**, v. 15, n. 7, p. 853–866, 2005.

OLSHAUSEN, B. A.; FIELD, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. **Nature**, v. 381, n. 6583, p. 607–609, 1996.

OLSHAUSEN, B. A.; FIELD, D. J. Sparse coding with an overcomplete basis set: A strategy employed by V1? **Vision Research**, v. 37, n. 23, p. 3311–3325, 1997.

OLSHAUSEN, B. A.; FIELD, D. J. Vision and the Coding of Natural Images: The human brain may hold the secrets to the best image-compression algorithms. **American Scientist**, v. 88, n. 3, p. 238–245, 2000.

PALMER, S. Fundamental aspects of cognitive representation. In: ROSCH, E.; LOYD, B. B. (Eds.). **Cognition and categorization**. Hillsdale NJ: Erlbaum, 1978.

PANOZ-BROWN, D. et al. Rats remember items in context using episodic memory. **Current Biology**, v. 26, n. 20, p. 2821–2826, 2016.

PENN, D. C.; HOLYOAK, K. J.; POVINELLI, D. J. Darwin's triumph: Explaining the uniqueness of the human mind without a deus ex machina. **Behavioral and Brain Sciences**, v. 31, n. 2, p. 153–178, 2008b.

PENN, D. C.; HOLYOAK, K. J.; POVINELLI, D. J. Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. **Behavioral and Brain Sciences**, v. 31, n. 2, p. 109–130, 2008a.

PERINI-SANTOS, E. Does the principle of compositionality explain productivity? For a pluralist view of the role of formal languages as models. **CEUR Proceedings of the Workshop on Contexts in Philosophy**, 2017.

PERKINS, D. N. The engine of folly. In: **Why smart people can be so stupid**. London: Yale University Press, 2002. p. 64–85.

PERLMAN, M. Pagan teleology: Adaptational role and the philosophy of mind. In: ARIEW, A.; CUMMINS, R.; PERLMAN, M. (Eds.). **Functions: New essays in the philosophy of psychology and biology**. [s.l.] Oxford University Press, 2002. p. 263–290.

PERNER, J. Understanding the representational mind. 1993.

PESSOA, L. **The entangled brain: How perception, cognition, and emotion are woven together**. [s.l.] The MIT Press, 2022.

PFEIFFER, B. E.; FOSTER, D. J. Hippocampal place-cell sequences depict future paths to remembered goals. **Nature**, v. 497, n. 7447, p. 74–79, 2013.

PHILLIPS, J. C.; WARD, R. S-r correspondence effects of irrelevant visual affordance: Time course and specificity of response activation. **Visual Cognition**, v. 9, n. 4-5, p. 540–558, 2002.

PICCININI, G. **Physical computation: A mechanistic account**. [s.l.] Oxford University Press, 2015.

PICCININI, G. **Neurocognitive mechanisms: Explaining biological cognition**. [s.l.] Oxford University Press, 2021.

PICCININI, G. Situated neural representations: Solving the problems of content. **Frontiers in Neurorobotics**, v. 16, 2022.

PINKER, S. **The language instinct**. New York: William Morrow; Company, 1994.

PINKER, S. **How the mind works**. [s.l.] Norton, 1997.

POVINELLI, D. **Folk physics for apes : The chimpanzee's theory of how the world works**. Oxford New York: Oxford University Press, 2000.

POVINELLI, D. **World without weight: Perspectives on an alien mind**. New York: Oxford, 2012.

POVINELLI, D. J.; BERING, J. M.; GIAMBRONE, S. Toward a science of other minds: Escaping the argument by analogy. **Cognitive Science**, v. 24, n. 3, p. 509–541, 2000.

PROUST, J. **The philosophy of metacognition**. Oxford: Oxford University Press, 2013.

PYLYSHYN, Z. W. **Computation and cognition**. [s.l.] MIT Press, 1984.

PYLYSHYN, Z. W. (ED.). **The robots dilemma: The frame problem in artificial intelligence**. [s.l.] Ablex, 1987.

RAMSEY, W. **Representation reconsidered**. Cambridge New York: Cambridge University Press, 2007.

RANSOM, M. Why emotions do not solve the frame problem. In: M{ULLER, V. (Ed.). **Fundamental issues of artificial intelligence**. [s.l.] Springer, 2016. p. 353–365.

RETT, A. et al. Children's use of causal structure when making similarity judgments. **Proceedings of the Annual Meeting of the Cognitive Science Society**, v. 43, 2021.

RIBEIRO, L. F. R.; SAVERESE, P. H. P.; FIGUEIREDO, D. R. **struc2vec: Learning node representations from structural identity**. Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, Aug. 2017.

RIDDOCH, M. J.; HUMPHREYS, G. W.; EDWARDS, M. G. Visual affordances and object selection. In: MONSELL, S.; DRIVER, J. (Eds.). **Control of cognitive processes: Attention and performance XVIII**. [s.l.] MIT Press, 2000. v. 18, p. 603–625.

RIETVELD, E. Context-switching and responsiveness to real relevance. In: KIVERSTEIN, J.; WHEELER, M. (Eds.). **Heidegger and cognitive science**. Hampshire: Palgrave Macmillan, 2012. p. 105–135.

ROLLA, G. Por que não somos só o nosso cérebro: Em defesa do enativismo. **TRANS/FORM/AÇÃO: Revista de filosofia**, n. 46, p. 207–236, 2023.

RUMELHART, D. E. et al. **Parallel distributed processing: Explorations in the microstructure of cognition: Foundations (volume 1)**. Cambridge, MA: MIT press, 1986. v. 1.

SÁ PEREIRA, R. H. DE; SOUZA FILHO, S. F. DE; BARCELLOS, V. M. Por que somos o nosso cérebro: O enativismo posto em questão. **TRANS/FORM/AÇÃO: Revista de filosofia**, n. 46, p. 517–554, 2023.

SALAY, N. **Why dreyfus' frame problem argument cannot justify anti-representational AI**. Proceedings of the annual meeting of the cognitive science society. v. 31, p. 1198–12032009.

SAMUELS, R. The complexity of cognition: Tractability arguments for massive modularity. In: CARRUTHERS, P.; LAURENCE, S.; STICH, S. (Eds.). **The innate mind: Structure and contents**. [s.l.] New York: Oxford University Press New York, 2005. p. 107.

SAMUELS, R. Classical computationalism and the many problems of cognitive relevance. **Studies in History and Philosophy of Science**, v. 41, n. 3, p. 280–293, 2010.

SCHANK, R. C.; ABELSON, R. P. **Scripts, plans, goals and understanding**. [s.l.] Taylor & Francis Inc, 1977.

SCHWITZGEBEL, E. A phenomenal, dispositional account of belief. **Noûs**, v. 2, n. 36, p. 249–275, 2002.

SCHWITZGEBEL, E. A dispositional approach to the attitudes. In: NOTTELMANN, N. (Ed.). **New essays on belief**. [s.l.] Palgrave, 2013. p. 75–99.

SHANAHAN, M. **Solving the frame problem: A mathematical investigation of the common sense law of inertia**. [s.l.] MIT, 1997.

SHEA, N. **Representation in cognitive science**. [s.l.] Oxford University Press, 2018.

SHIN, S.-J. **The logical status of diagrams**. [s.l.] Cambridge University Press, 1994.

SIMON, H. **The sciences of the artificial**. Cambridge, Mass: MIT Press, 1996.

SMITH, B. C. **The promisse of artificial intelligence: Reckoning and judgment**. London, England: MIT Press, 2019.

SMOLENSKY, P. Tensor product variable binding and the representation of symbolic structures in connectionist systems. **Artificial Intelligence**, v. 46, n. 1-2, p. 159–216, 1990.

SMOLENSKY, P.; LEGENDRE, G.; MIYATA, Y. **Principles for an integrated connectionist/symbolic theory of higher cognition**. [s.l.] Institute of Cognitive Science - University of Colorado, 1992.

SPERBER, D. Modularity and relevance: How can a massively modular mind be flexible and context-sensitive. In: CARRUTHERS, P.; LAURENCE, S.; STICH, S. (Eds.). **The innate mind: Structure and contents**. [s.l.] Oxford University Press, 2005. p. 53–68.

SPERBER, D.; WILSON, D. **Relevance: Communication and cognition**. [s.l.] John Wiley; Sons, 1995.

SPERBER, D.; WILSON, D. Fodor's frame problem and relevance theory. **Behavioral and brain sciences**, v. 19, n. 3, p. 530–532, 1996.

STALNAKER, R. **Inquiry**. [s.l.] MIT Press, 1984.

STERKENBURG, T. F.; GRÜNWALD, P. D. The no-free-lunch theorems of supervised learning. **Synthese**, v. 199, n. 3-4, p. 9979–10015, 2021.

SWOYER, C. Structural representation and surrogative reasoning. **Synthese**, v. 87, n. 3, p. 449–508, 1991.

TENNIE, C.; CALL, J.; TOMASELLO, M. Ratcheting up the ratchet: On the evolution of cumulative culture. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 364, n. 1528, p. 2405–2415, 2009.

THELEN, E.; SMITH, L. **A dynamic systems approach to the development of cognition and action**. Cambridge, MA: MIT Press, 1992.

THOMPSON, R. K. R.; ODEN, D. L.; BOYSEN, S. T. Language-naive chimpanzees (pan troglodytes) judge relations between relations in a conceptual matching-to-sample task. **Journal of Experimental Psychology: Animal Behavior Processes**, v. 23, n. 1, p. 31–43, 1997.

THORUP, K. et al. Evidence for a navigational map stretching across the continental u.s. In a migratory songbird. **Proceedings of the National Academy of Sciences**, v. 104, n. 46, p. 18115–18119, 2007.

TODES, S. **Body and world**. Cambridge: MIT Press, 2001.

TURING, A. M. On computable numbers, with an application to the entscheidungsproblem. **Proceedings of the London mathematical society**, v. 2, n. 1, p. 230–265, 1936.

UZGIRIS, I. C. et al. Contextual influences on imitatives interactions between mothers and infants. In: LOCKMAN, J. J.; HAZEN, N. L. (Eds.). **Action in social context: Perspectives on early development**. [s.l.] Springer, 1989. p. 103–127.

VAN GELDER, T. What might cognition be, if not computation? **The Journal of Philosophy**, v. 92, 1995.

VARELA, F. J.; ROSCH, E.; THOMPSON, E. T. **The embodied mind: Cognitive science and human experience**. [s.l.] MIT Press, 1991.

VERVAEKE, J. Rationality and relevance realization. **Forthcoming**, 2022.

VERVAEKE, J.; FERRARO, L. Relevance realization and the neurodynamics and neuroconnectivity of general intelligence. In: **SmartData**. [s.l.] Springer New York, 2013. p. 57–68.

VERVAEKE, J.; LILLICRAP, T. P.; RICHARDS, B. A. Relevance realization and the emerging framework in cognitive science. **Journal of Logic and Computation**, v. 22, n. 1, p. 79–99, 2012.

VON UEXKULL, J. A stroll through the worlds on animals and men. In: LASHLEY, K. (Ed.). **Instintictive behavior**. [s.l.] International Universities Press, 1934.

WAAL, F. B. M. DE. Intentional deception in primates. **Evolutionary Anthropology: Issues, News, and Reviews**, v. 1, n. 3, p. 86–92, 1992.

WAAL, F. DE. **Chimpanzee politics: Power and sex among apes**. [s.l.] Johns Hopkins University Press, 2007.

WALKER, C. M.; GOPNIK, A.; GANEA, P. A. Learning to learn from stories: Children's developing sensitivity to the causal structure of fictional worlds. **Child Development**, v. 86, n. 1, p. 310–318, 2014.

WASKAN, J. A. Intrinsic cognitive models. **Cognitive Science (Elsevier Science)**, v. 27, 2003.

WASKAN, J. A. **Models and cognition: Prediction and explanation in everyday life and in science**. Cambridge: MIT Press, 2006.

WEBB, B. Modeling biological behaviour or dumb animals and stupid robots. **Pre-Proceedings of the Second European Conference on Artificial Life**, p. 1090–1103, 1993.

WHEELER, M. **Reconstructing the cognitive world: The next step (MIT press)**. [s.l.] A Bradford Book, 2005.

WHEELER, M. Cognition in context: Phenomenology, situated robotics and the frame problem. **International journal of philosophical studies**, v. 16, n. 3, p. 323–349, 2008.

WHEELER, M. Naturalizing dasein and other (alleged) heresies. In: KIVERSTEIN, J.; WHEELER, M. (Eds.). **Heidegger and cognitive science**. Hampshire: Palgrave Macmillan, 2012. p. 176–212.

WITTGENSTEIN, L. **Tractatus logico-philosophicus**. Translation: C. K. Ogden. New York: Harcourt, Brace & Company, 1922.

WOLPERT, D. H.; MACREADY, W. G. No free lunch theorems for optimization. **IEEE Transactions on Evolutionary Computation**, v. 1, n. 1, p. 67–82, 1997.