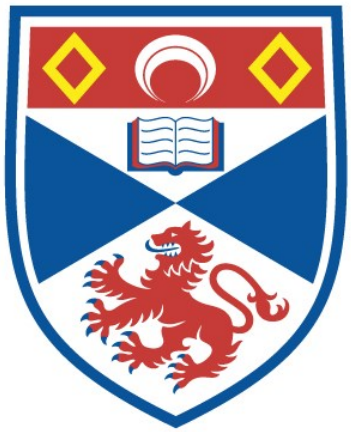
The mark of the cognitive

Niccolò Aimone Pisano

A thesis submitted for the degree of PhD
at the
University of St Andrews



2024

Candidate's declaration

I, Niccolò Aimone Pisano, do hereby certify that this thesis, submitted for the degree of PhD, which is approximately 73,989 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for any degree. I confirm that any appendices included in my thesis contain only material permitted by the 'Assessment of Postgraduate Research Students' policy.

I was admitted as a research student at the University of St Andrews in September 2019.

I confirm that no funding was received for this work.

Date 19th June 2023 Signature of candidate

Supervisor's declaration

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree. I confirm that any appendices included in the thesis contain only material permitted by the 'Assessment of Postgraduate Research Students' policy.

Date 19th June 2023 Signature of supervisor

Permission for publication

In submitting this thesis to the University of St Andrews we understand that we are giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. We also understand, unless exempt by an award of an embargo as requested below, that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that this thesis will be electronically accessible for personal or research use and that the library has the right to migrate this thesis into new electronic forms as required to ensure continued access to the thesis.

I, Niccolò Aimone Pisano, confirm that my thesis does not contain any third-party material that requires copyright clearance.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

Printed copy

Embargo on all of print copy for a period of 1 year on the following ground(s):

- Publication would preclude future publication

Supporting statement for printed embargo request

I intend to use some of the material coming from a few chapters for the preparation of one or two papers for publication.

Electronic copy

Embargo on all of electronic copy for a period of 1 year on the following ground(s):

- Publication would preclude future publication

Supporting statement for electronic embargo request

I intend to use parts of this thesis to prepare one or two papers for publication

Title and Abstract

- I agree to the title and abstract being published.

Date 19th June 2023 Signature of candidate

Date 19th June 2023 Signature of supervisor

Underpinning Research Data or Digital Outputs

Candidate's declaration

I, Niccolò Aimone Pisano, hereby certify that no requirements to deposit original research data or digital outputs apply to this thesis and that, where appropriate, secondary data used have been referenced in the full text of my thesis.

Date 19th June 2023 Signature of candidate

Abstract

Several philosophical debates in the philosophy of mind and of the cognitive sciences seem to require the elaboration of a mark of the cognitive (MOC). Some proposals for individually necessary and/or jointly sufficient conditions for cognition are already available, but each of them is not entirely satisfactory for different reasons.

I start by drawing on some of the extant proposals, and I advance a possible candidate necessary condition for cognition. I motivate the claim that cognition requires the existence of a first-person perspective (1PP) associated with it. However, while reasonable, I argue that we should not ultimately accept this claim. Moreover, some of the reasons that I provide for not accepting it also apply to the broader methodological family of approaches to cognition that has been labelled by Lyon (2006) the "anthropogenic" family. As a result, it is not just the 1PP condition that needs to be dismissed, it is the entire anthropogenic approach to cognition that should not be pursued in attempting to elaborate a MOC.

Luckily, the other broad methodological family, the one of the "biogenic" approaches to cognition, is not vulnerable to the same issues that arise in the case of anthropogenic approaches. We should therefore adopt a biogenic approach to the issue of finding a mark of the cognitive. Nevertheless, elaborating a MOC within a biogenic framework may not prove as beneficial as one may hope. In fact, what appears to be the most promising and positively received product of a biogenic approach, namely the Free-Energy Principle and the accounts of cognition based on it, needs to be understood in instrumentalist terms. Consequently, while we may still be able to achieve an understanding of what cognition is, some of the debates meant to be settled by the elaboration of a MOC may remain unsettled.

Table of Contents

Introduction

A widely agreed upon bit of philosophical wisdom is that a first step towards the resolution of philosophical disputes consists in avoiding talking past each other. This is why it is customary in analytic philosophy to always clarify as carefully as possible the terminology one employs in their works. But, of course, tidying the terminology up will not settle every disagreement; the substantial ones will remain.

Sometimes genuine philosophical disagreements look rather similar to merely verbal disputes. Just like in verbal disputes the issue is overcoming the parties' idiosyncratic conceptual differences in order to reach the core of the issue, some genuine disagreements require finding a solid conceptual common ground to start building a solution to the philosophical issue (or issues) in question. In this thesis, I will be concerned with one such genuine disagreement, namely that over the nature of cognition.

When it comes to studying "the nature" of some natural phenomenon, there are many ways to approach the subject. One may try to find the essence of that natural phenomenon, that is, very roughly, what that phenomenon intrinsically and necessarily is. Or perhaps one may just be interested in formulating a theory of how that natural phenomenon works. Or yet again, one may try to find a set of necessary and/or jointly sufficient conditions for that natural phenomenon to occur, or, to put it differently, for something to qualify as an instance of that phenomenon. Here, I will follow what is by now the traditional approach in the philosophy of the cognitive sciences, and I will explore the nature of the condition from this third angle. That is to say, I will discuss the issue of finding a set of individually necessary and/or jointly sufficient conditions for cognition to occur.

The idea of undertaking such a task is not new by any means. Indeed, the thought that in order to resolve certain philosophical disputes related to the 4E (embodied, embedded, extended, enactive) views on cognition we need what is now commonly called a "mark of the cognitive" has been around for the past quarter of a century. Despite there being plenty of scepticism concerning the viability or even the usefulness of the project (as reported by Adams (2010)), the issue of finding a mark of the cognitive stimulated the

production of a significant amount of literature. So, why should we be interested, and why have many been interested, in finding a mark of the cognitive?

The answer to this question has to do with what a mark of the cognitive is meant to do for us. First and foremost, the purpose of a mark of the cognitive (MOC, hereafter) is to demarcate the domain of the cognitive from that of the non-cognitive. The background assumption for this *desideratum* for a MOC is that not everything in the world should be granted cognitive status, for the simple reason that not everything in the world manifests cognition. This seems uncontroversial unless one is sympathetic towards some form of panpsychism, which I will not discuss in the present work, although I will touch upon analogous issues in a number of places. But then, of course, one might ask why it is important to be able to tell whether something counts as a manifestation of cognition or not. In a way, explaining exactly why it matters to know whether some phenomenon is cognitive in nature presupposes some knowledge of what cognition is, for if one does not know what it means to say that something counts as a cognitive process, one cannot know what features cognitive processes possess that make this claim interesting. For this reason, that is obviously not an explanation that can be given before already having a MOC at one's disposal. Hence, the motivation for wanting to discriminate between what is cognitive and what is not cannot *ex ante* depend on the content of the MOC one will ultimately come up with. Rather, it has to do with what one hopes to be able to do with such knowledge. In particular, knowing whether certain phenomena are cognitive or not allows one to narrow down the domain of the discussion for other interesting questions. To mention a few among the currently discussed ones: are AI's capable of thought? How will we tell if they will ever be able to think? Are plants capable of cognition? If so, are they capable of thought in a way that warrants giving them rights, just like animals have rights?

In short, the main reason why it would be desirable to find a MOC is that knowing what cognition is would provide a solid basis to investigate other, important philosophical questions (in addition to being interesting in its own right). However, many would here urge caution. While it is true that a number of disputes over philosophical questions would benefit from the formulation of a MOC, it is not clear what impact the elaboration of a MOC would have on the practice of the cognitive sciences. How could a neurophysiologist make use of a theoretical account of the overall nature of the

phenomena they study? Would a psychiatrist stop treating certain disorders if it turned out that they are not cognitive disorders? Should psychologists appropriate the study of plants, if they turned out to be capable of cognition, or should they leave it to botanists? Even though these are interesting questions, they do generally bring to the fore the idea that actual scientific practice is unlikely to be significantly affected by the ability to distinguish in a principled way what is cognitive from what is not. Nonetheless, there are ways in which scientific practice could be affected by such ability. For instance, if one is interested in describing how cognitive processes work, it is important to be able to draw a line between the systems under study and their environment, in order to both avoid excluding important portions of the relevant system and including unimportant and side-tracking bits.

This brings us to another central reason why we should be interested in seeking a MOC. Not only would a MOC allow us to distinguish what is cognitive from what is not, but it would also allow us to demarcate the boundaries of cognitive systems. This issue has become one of the most pressing ones in the aftermath of one of the four "E's" of the 4E's views: the extended cognition view. The core idea of this view (originally introduced by Clark and Chalmers (1998)) is that entities and processes situated outside the biological boundaries of the brain, or even of the body, are constitutive of the cognitive processes that we instantiate. If this claim were true, not only would we be forced to reconceptualise who we are as thinkers, and how we work as such, but it would also have other interesting implications. For instance, it may be the case that some people with certain neural impairments would not be considered cognitively impaired anymore, if, using external tools, they are able to perform the relevant cognitive tasks in a similar way to people without those neural impairments.

The extended cognition view is notoriously controversial, but it is also the crucible of the mark of the cognitive project. It is in fact in order to argue against this view that some of the first explicit attempts at characterising cognition in recent years were made (Adams and Aizawa (2001, 2008)). The idea was that, despite the *prima facie* implausibility of the extended cognition view, one needs more substantial reasons to reject it (provided that one wishes to do so, that is). Hence, the strategy adopted by Adams and Aizawa was to show that the reason why the extended cognition view is false of actual, currently existing cognitive systems, is that the external constituents of the

allegedly extended cognitive systems are not really constitutive of the relevant cognitive processes, but merely causally coupled with them. In fact, they provided a criterion on the basis of which to distinguish the domain of the cognitive from that of the non-cognitive, and, on the basis of this MOC, they maintained that it is possible to demarcate the boundaries of cognitive systems in a way that falsifies the extended cognition view[1].

The debate over the extended cognition view thus led to the birth of the mark of the cognitive proposal, and it did so in a way that reflects the third aforementioned way to account for the "nature" of something. Namely, the approach to the issue that became the standard one adopted in the search for a MOC consists not in trying to capture the essence of cognition, nor to focus (solely) on how cognition works, but to formulate a set of necessary and/or sufficient conditions for the occurrence of cognition: all cognitive systems must meet such conditions, and/or any system that meets them qualifies as cognitive. However, this raises another issue. One thing is being able to draw the boundaries of a cognitive system, another is establishing that all that happens within such boundaries, and that meets the conditions for cognition, is part of the cognitive economy of one and the same cognitive subject. In fact, it may well be the case that two spatially co-located cognitive processes should not be ascribed to one and the same entity, and, conversely, that two processes should instead be ascribed to the same entity. Consider, for example, mental disorders involving dissociative episodes to some degree. Perhaps one may be intuitively inclined to say that cognitive processes happening during dissociative episodes are not ascribable to the person undergoing the episodes in the same way as the person's ordinary cognitive processes are ascribable to that person. Or perhaps one may have the opposite intuition and be inclined to say that they are ascribable to that person, just like ordinary cognitive processes are. Either way, it is clear that just knowing that the processes in question are cognitive and that they are spatially occurring within the same organism does not tell us much about whether they should be regarded as part of the cognitive economy of one and the same cognitive subject. This is therefore another job that we should expect a good MOC to do for us: enabling us to tell, in a principled way, whether some cognitive process belongs to some cognitive subject or not.

[1] Their proposal, in a nutshell, is that cognition must involve nonderived content. I will discuss Adams's and Aizawa's view extensively in the first chapter.

To summarise, there are many reasons why we should be interested in finding a MOC. First, knowing what cognition is, in the sense of knowing what conditions are to be met for cognition to occur, allows us to distinguish cognitive from non-cognitive phenomena. This, in turn, can be helpful in the discussion of other, cognition-related philosophical questions. Second, a MOC would allow us to demarcate the boundaries of cognitive systems, which is useful in both philosophical and scientific[2] contexts. Finally, a MOC can (or should) give us principled criteria to ascribe certain cognitive phenomena to a given cognitive subject; that is to say, it can shed light on what it means for a cognitive process to belong to a cognitive subject.

As I have mentioned, several proposals for a MOC have been formulated over the years. In this thesis, I will kick off the discussion by analysing and commenting on some of them in order to highlight their strengths and weaknesses. Then, drawing from the results of that discussion, I will advance a possible necessary condition for cognition, namely that there must be a first-personal dimension associated with putatively cognitive processes. I will extensively motivate and eviscerate it, not only because it is a plausible one, but also because its discussion instructively brings to the fore the core issues related to the MOC project. However, I will argue that this "first-person-perspective condition" should not be endorsed. Specifically, an important reason why it should not be endorsed is that it is the product of a wrongheaded approach to the search for a MOC, namely the anthropogenic approach, which takes human cognisers as the starting point of the inquiry. I will thus recommend adopting a different, alternative approach, that is, the biogenic approach, which instead takes biological organisms as its starting point. I will conclude by showing what a biogenic approach to the issue of finding a MOC would look like, focusing my attention on the Free-Energy Principle, which in my opinion is the view that most clearly engenders the biogenic approach[3].

Here is how I will proceed, in more detail. In Chapter 1 I will present and comment on three extant proposals for a mark of the cognitive. I will start by analysing Adams's and Aizawa's (2001, 2008) claim that nonderived content is necessary for cognition. I will suggest that we should understand nonderived content not only as that content which

[2] For instance, if cognition is tied to the computation of some well-defined quantity (e.g. free-energy), being able to demarcate the boundaries of cognitive systems would prove extremely valuable in scientific contexts.

[3] Once paired with some process theory, that is. More on that in the last chapter.

is possessed independently from the influence of any pre-existing form of intentionality, but also, and crucially, as that kind of content which is immediately accessible. I will then show that this idea that there is a component of immediacy involved in cognitive processing applies also to the two other proposals for a MOC I will discuss in this chapter, namely Adams's and Garrison's (2013) claim that cognition needs to be related to the cogniser's own reasons, and Rowlands's (2009, 2010) set of jointly sufficient conditions for cognition. In particular, the ability to gain immediate access to the information processed in cognition is a distinguishing characteristic of what a cognitive subject is.

In Chapter 2 I will reach similar conclusions to the ones reached in the previous chapter, but via a different route. I will start arguing that what materially realises cognition can hardly be non-vaguely identified, and I will accordingly consider a comprehensive candidate necessary condition for cognition, namely the claim that cognition must involve the processing of information. After having shown that this is indeed not a sufficient characterisation of cognition, I will argue, on predominantly epistemic grounds, as opposed to metaphysical grounds, that a satisfactory account of cognition should not neglect the first-personal component associated with it. In doing so, I will link the idea of a first-person perspective (1PP) to that of a cognitive subject.

In Chapter 3, I will motivate the claim that the 1PP may be necessary for cognition. The 1PP is a non-thematizing perspective, which has a tight connection with the notion of a cognitive subject, and by means of which the information contained in cognitive states or processes is made available. Differently from the previous chapter, I will develop a metaphysical argument, rather than an epistemic one. That is to say, I will not defend this claim because failing to endorse it would make cognition unintelligible, but because, in the absence of a 1PP associated with it, cognition would be disqualified as a real, natural phenomenon. The core reason is that, were there not a 1PP associated with cognition, an element of arbitrariness would enter the picture: cognition would be characterised by too weak features that would not suffice to determine the boundaries of cognitive systems. But the arbitrariness that would be consequently involved in the determination of such boundaries is not acceptable for a genuine natural phenomenon.

In Chapter 4 I will begin my criticism of the claim that the 1PP is necessary for cognition. Despite the plausibility of this claim, there are reasons to be suspicious of it. In particular, it seems that there are non-negligible connections between the 1PP and

consciousness. If a substantial link between the two is indeed in place, then it follows that, were one to endorse the necessity of the 1PP for cognition, one risks being committed to the further claim that there can be no unconscious cognitive processes. This goes against scientific consensus, but *per se*, going against scientific consensus is not damning. Most of the fourth chapter will thus be devoted to the explanation of how, in this specific case, going against scientific consensus is undesirable. In particular, I will show that the absence of a theory of consciousness able to support the 1PP condition is one of the main reasons why departure from consensus is damning in this context. This will be done by showing how the Integrated Information Theory of consciousness, which is plausibly the most suitable available theory of consciousness for the purpose of defending the 1PP condition, is not sufficiently well-developed to do the job.

In Chapter 5 I will elaborate a further argument against the 1PP condition on cognition. However, the scope of this argument is much broader than that of the argument presented in the previous chapter. In fact, the core idea is that all anthropogenic accounts of cognition, and the 1PP condition as a member of that category, are not viable. My argumentative strategy will proceed by dilemma: either one is nonreductionist, or one is reductionist. If the former, then the causal exclusion problem makes the properties appearing in one's account of cognition epiphenomenal. This is unacceptable in the light of Alexander's Dictum, which states that something can only be real and potentially characterise a natural phenomenon (in this case, cognition) if it possesses causal properties. One must therefore couch one's account of cognition in a reductionist conceptual framework. But this is not a move that pursuers of anthropogenic accounts of cognition can (at least at present) make. However, one does not encounter analogous problems in adopting a biogenic, instead of an anthropogenic, approach to the issue of finding a mark of the cognitive. The conclusion of this chapter will therefore be that we should pursue a biogenic approach.

In Chapter 6 I will examine what is arguably the currently most significant way to pursue a biogenic approach to cognition, namely the Free-Energy Principle (FEP). I will present and analyse the FEP, and I will argue that, while most of its defenders are realists about both the FEP itself and the models of cognitive systems that can be elaborated within its conceptual and mathematical framework, we should conceive of the models of cognitive systems constructed under the FEP in instrumentalist terms. This, however,

does not prevent us from gaining precious insights into the nature of cognition. It is in fact possible to elaborate a MOC within the FEP framework, as the existence of various such proposals shows. Nevertheless, instrumentalism comes at a price. That is, if we have an instrumentalist conception of our models of cognitive systems, it is unlikely that we will be able to settle several (but not all) of the debates from which the interest in elaborating a MOC stemmed.

Chapter 1

Introduction

In this chapter, I will begin the quest for the elaboration of a mark of the cognitive (MOC) by examining three of the extant proposals. The first is a necessary condition advanced by Adams and Aizawa (2001, 2008), according to whom cognitive processes need to involve, at least to some extent, the manipulation of representations endowed with nonderived content. The second account is another necessary condition, argued for by Adams and Garrison (2013), based on which the performance of cognitive processes needs to take place for reasons internal to the system at issue. Finally, I will present Rowlands's (2009, 2010) set of jointly sufficient conditions.

The main purpose of this chapter is to get things started by reviewing some attempts that have already been made to characterise cognition. However, this will not be a purely contextualising exercise, and my commentary on the accounts discussed here will prepare the grounds for the discussion of the distinction between anthropogenic and biogenic approaches to cognition (Lyon (2006)) that will take place in the fifth chapter. This is because all three accounts fall, arguably, within the anthropogenic camp, and are indeed good representatives of it, while my favoured account of cognition falls within the other camp.

I will proceed as follows. In the first section, I will present and comment on Adams's and Aizawa's proposal. Then, in the second section, I will more extensively discuss the notion of nonderived content claimed by the authors to play an important role in determining whether the cognitive status of putatively cognitive processes should indeed be granted or not. In doing so, I will argue in favour of an understanding of nonderived content not merely as content which does not require the pre-existence (and activity) of some intentional agent, but as content which is immediately accessible to its user. In the third section I will draw upon the preceding considerations, and I will argue that the occurrence of nonderived content seems to require the simultaneous existence (but not the pre-existence) of a subject.

Finally, in the fourth and fifth sections, I will present and comment upon the two other proposals of MOC's, namely Adams's and Garrison's one, and Rowlands's one. This will be done by making use of the ideas developed in the previous sections; in particular, I will comment upon the relevance of the role that the notion of subjectivity, in the sense of "being relative to a subject", appears to play in these accounts too.

1. Adams' and Aizawa's view

In their works devoted to the criticism of the Extended Mind view on cognition, Adams and Aizawa (2001, 2008, 2010a, 2010b) have suggested a necessary condition to be met by a state or process to qualify as cognitive. Namely, they have argued that a crucial component of a MOC is that the content of the states and processes involved in cognitive phenomena is nonderived. Although nonderived content as it appears in Adams's and Aizawa's (AA) works can be given a fairly straightforward definition as the intentional content which something possesses independently from any meaning-conferring act performed by other intentional entities, not much is offered in their works by way of explanation of this notion. In particular, it is not very clear what the specific contribution nonderived content would provide to the instantiation of a cognitive process would be, which derived content cannot offer. This is problematic if one claims that nonderived content is necessary for cognition, as there would not be adequate grounds to explain *why* that would be the case.

The purpose of this section is to analyse the notion of nonderived content as it appears in AA's proposed MOC. Taking my moves from it, in the next section I will argue that the role that it is supposed to play in leading to the occurrence of cognitive processes does not stem as much from its being independent of some prior intentionality, as it instead does from its characterisation as content which is immediately available to the system possessing it.

The idea behind the notion of nonderived content, also referred to as "original intentionality" or "intrinsic content", is that not all intentional states or processes possess their content in virtue of the influence of some prior intentional phenomenon. This is because, to avoid infinite regresses or vicious circles, it seems that at least some intentional phenomena should not be dependent upon other intentional facts for their

existence as such. Accordingly, there have been many attempts at naturalising intentionality, i.e. attempts at grounding some presumably primitive forms of intentionality, such as those involved in mental activities, upon non-intentional facts[4]. However, such attempts have not so far proved uncontroversially successful and, according to some (Kriegel (2011, pp.3-4)), they are to be considered expressions of a degenerative research project. Nonetheless, the intuition that there must be intentional phenomena whose content does not depend upon other intentional phenomena remains untouched.

It is this sort of intentional phenomena that AA have in mind in talking of states or processes with nonderived content (be they representations or not, taking part in a language of thought or not, as Adams and Aizawa (2001, p.49) remark). According to them, it is evident that, differently from other prototypical cases of intentional entities (for instance, words of natural languages), cognitive states cannot rely upon external factors such as stipulations made by intentional systems or social practices[5] for their possession of the meanings they have. For example, there is nothing in the sequence of three letters "c-a-t", neither as a type nor as a token, however instantiated it might be (spoken or written), that entertains a special relation with cats independently from the existence of a linguistic community establishing so. This is the reason why, for instance, if someone encounters a word from an unknown language it will not be possible to gain access to its meaning solely in virtue of the formal properties of the word taken in isolation from the rest of the relevant language-related facts. Furthermore, as a matter of fact, many words have changed their meanings over time without all of their recent users being aware of the older meanings, and this is additional clear evidence in support of the fact that there is no intrinsic connection between a word and its meaning. What matters is what such relation is taken to be by entities already independently involving some forms of intentionality, like humans.

I do not want to delve into the discussion of whether and to what extent the specific character of other sorts of intentional entities such as pictures or artworks in general depends upon the prior existence of some form of intentionality. What matters here is

[4] Some notable proposals have been put forward by Fodor (1987, 1990), Millikan (1984), and Dretske (1981).

[5] It does not matter here whether we should think of communities as collective intentional subjects, or as the mere sum of individual intentional subjects.

AA's view that the existence of intentional phenomena would be rather mysterious if there were not at least some form of primitive intentionality, namely if there were not at least some kind of intentional entity which is able to refer without this being the result of a more or less arbitrarily determined external influence exerted by some pre-existing intentional agent, regardless of the sort of influence that is exerted. This primitive, original intentionality is the kind of intentionality that AA have in mind when they talk about intentional entities having nonderived content. Originally intentional entities have the content they have independently from external attributions, conventions, social practices and so on: they do not derive their content from any prior intentionality.

As it happens, AA's intuition concerning original intentionality is not universally shared, as Dennett's (1994) argument against its existence famously testifies. His argument is developed from a thought experiment in which a person builds a machine capable of interacting, for centuries, as intelligently and quickly with its environment as a regular human cogniser would do. This is done in the absence of human supervision, for the purpose of preserving the frozen, but still alive, body of the person until their awakening in a distant future. According to Dennett, the internal states of the robot only have derived intentionality, in that the robot has been programmed and assembled by a human creator. More specifically, even if the robot undergoes states and processes which had not, and could not have been, predicted in advance by its creator, such states and processes would still have just derived content, because their intentionality ultimately depends on the fact that the creator built the robot in the way they did. Furthermore, the same would hold for any additional robot that the original one itself might happen to assemble at some point in time, for those successive robots' intentionalities would come into being only as a product of the manifestations of the first robot's previously existing intentionality. That is to say, non-original intentionality is (at least sometimes) possessed transitively: if a content-bearing structure *X* comes to possess intentional content as the result of *Y*'s action, and the representations possessed by *Y* and involved in *Y*'s action all have derived content, then *X*'s content will plausibly be derived.

Next, Dennett moves a step further and argues that, if what just noted about the derivative character of the robot's intentional states is correct, then a similar line of reasoning would apply to humans as well, which are instead taken by supporters of original intentionality as the paradigmatic case of entities with nonderived content. After

all, if the human mind works the way it does, it is in virtue of the way it ultimately originated from the human genes (together with the interactions of the organism with the environment), which are passed from one generation to the following. Therefore, the presumed original intentionality displayed by any human individual really is not original in Dennett's view, because it depends on how human genes determine it to be. This means that even prototypical cases of original intentionality should be understood, according to Dennett, as instances of derived intentionality.

Adams and Aizawa (2005) have observed, in my opinion correctly, that Dennett's line of reasoning is flawed, in that it conflates two different sorts of determination relation with respect to intentionality. On the one hand, the existence of an intentional entity can be causally determined by some external, non-intrinsically intentional factor. This is the case of human minds, which are brought about entirely causally, if indirectly, by human genes, which in themselves, leaving aside metaphorical interpretations taking them to have intentional content, do not possess any intentional content. On the other hand, an intentional entity can be instantiated as such in virtue of a semantic, meaning-conferring relation, in addition to a causal one. This is the case, by assumption, of the robot in Dennett's thought experiment, whose intentional states are indirectly determined by its human creator in such a way that meanings are conferred to its internal states.

It seems, then, that what distinguishes the way in which intentional entities come to have derived content is the fact that the determination relation(s) is (are) not only causal but also meaning-conferring. If this is correct, then we are in the position to account for the existence of forms of intentionality which are grounded upon something other than intentional entities, namely exclusively on causally relevant facts. That is, according to AA, an intentional entity has nonderived content just in case it has the content it has exclusively in virtue of meeting the requirements set by a (naturalistic) account of intentionality.

For present purposes, it is not important to outline the sort of naturalised semantics that AA might want to appeal to. After all, they explicitly (2010b, p.588) do not commit themselves to any specific account of this sort (among the alternatives to Fodor's asymmetric causal dependency theory mentioned by AA are Cummins's (1996) isomorphism-based and Dretske's (1988) function of indications theories). For this reason, I will leave aside the issue of whether the naturalistic views on semantics they

presuppose are complete and/or correct or not, or even if there is an acceptable one at all among the available ones. Rather, I will just say that, up to this point, I agree with AA's stance on what nonderived content is. However, I think that, if this were the whole story about it, it would not be clear why the instantiation of nonderived content would be a necessary condition for the occurrence of cognition. We may grant that nonderived content is possessed solely in virtue of meeting some naturalistic set of requirements, while derived content is possessed in virtue of a meaning-conferring act. Nonetheless, this "etiological" account would not amount to anything more than a description of how two different forms of intentionality can make their appearance. It would not tell us why the former (original intentionality) but not the latter (derived intentionality) might be taken to be necessary for cognition, in that it is not clear what the specific contribution of nonderived content is. In the next section, I will elaborate on this point.

2. Nonderived content

2.1 Nonderived intentionality as a relative notion

Recall that nonderived content can be understood at first in terms of its counterpart, derived content. According to Adams and Aizawa (AA), a representation has derived content if it is bestowed with the content it has by some external intentional agent; otherwise, it has nonderived content. Words are paradigmatic examples of representations with derived content, in that they possess their contents not in virtue of their very nature, but because it is conventionally accepted (explicitly or implicitly; this is not relevant for the present purposes) that they do. For instance, the word "cat" refers to cats because in English it is conventionally attributed such meaning, but there is nothing in the word in itself, nor in any occurrence of it, that establishes a link between such representation and its content. This point is corroborated by the fact that a person who does not know English at all cannot gain access to the meaning of the word "cat", if exposed to an utterance of it (be it in oral or in written form).

By contrast, representations with nonderived content are such because they meet the requirements set by an account of original intentionality. In addition to that, I contend that their content is somehow immediately made available to their owners simply as a result of the owner's being presented with them. That is to say, since it is not a matter of

convention that a representation of this kind has the content it has, no interpretation is needed to access the content: it is not necessary to be able to apply an interpretation function to move from the formal component of the representation to its semantic component. All that is required is that the subject engaging with a representation of this kind is capable of accessing its content simply in virtue of the presentation of such representation[6]. Although this is not what makes the case that nonderived content arises, it seems to me that this point straightforwardly follows from the core characterisation of nonderived content. In a way, the immediate accessibility of nonderived content is as distinctive a feature of original intentionality as the satisfaction of the requirements set by an account of original intentionality is.

This characterisation of the notion of nonderived content already foreshadows my worries concerning it. In fact, it is evident to me that if the content of a representation[7] is not accessed by means of a more or less explicitly formulated conventionally established rule, then the successful access to that representation's content depends equally on the properties of the representation and on the inherent predisposition of the subject engaging with it to immediately access the content of a representation of that sort. But, since AA's condition on nonderived content is taken to be necessary for cognition, it seems that there are only two possible alternatives. Either (1) the notion of nonderived content is understood as a relative one, so that a representation can have nonderived content for some cognisers, and at best derived content for other, differently constituted ones. Or (2) cognisers must all have the same "interpretive apparatus", that is, they all must gain access to the content of representations with nonderived content in the same way, so that representations with nonderived content are such for every possible cogniser. The latter option would prevent, for example, inorganic machines from being cognisers. This is for the simple reason that, if the human brain states possess nonderived content, given that an inorganic machine would need some mediation in order to have access to their contents, the way machines gain access to those paradigmatic representations with nonderived content would be different from the way humans do. As a consequence, since

[6] Notice that this is compatible with semantic externalism. Semantic externalism concerns the way content-bearing representations originally come to have a content, but my point here is about how such content later becomes a part of a subject's cognitive economy.

[7] Adams and Aizawa do not argue in the favour of the representationalist claim that nonderived content must be possessed by representations, but for simplicity I will refer to the cognitive states having such content as representations.

they would not be able to make use of nonderived content in the way cognisers ought to, they would be disqualified as cognisers. In what follows, I will elaborate on these two alternatives.

Clearly, the second horn of the dilemma goes against many of the tenets of contemporary studies in cognition. Most importantly, it would not only be a clear version of biological (human) chauvinism, but it would also constitute a neat violation of functionalist multiple realisability. This is because, if the access to the content of a representation does not depend on the application of some previously acquired interpretive method[8], but rather it takes place immediately (in the literal sense of "without mediation") because the representation in itself makes its own content available to the cogniser, then this process can only happen in virtue of the physically identifiable properties of the representation in question. If this is the case, then differently constituted cognisers, such as animals of different species, would not be able to gain access in the same immediate way to the content of a representation which is endowed with nonderived content for other cognisers, such as humans. This is for the same reason I have mentioned above relatively to machines. Suppose that a cognitive system is realised differently from another, so that, say, one is a human brain and the other a set of electronic circuits. Then, a particular representation with nonderived content in the first system would, at best, have derived content in the latter case, because it would need a process of interpretation in order to be made accessible, with its content, to the electronic cognitive system. The reason for this is that if the non-derivativity of the content of a representation depends, as it seems to me, on the properties of its particular occurrences, then it would be impossible to have immediate access to the content of some representation, for those cognisers whose underlying physical processes are of a different nature from the nature of the processes relevant for the particular tokens of the type of representation in question, occurring within a system where such occurrences have nonderived content.

Suppose that whenever I think of cats I am in a certain neurological configuration, call it *A*. Then, since I am a cogniser, if the condition on nonderived content is a necessary one, *A* must somehow be (associated with) a representation with nonderived content, that is, the content of *A* must be accessible to me without the need for me to interpret it. *A*, or

[8] The cogniser does not need to be explicitly and thoroughly aware of such a procedure, as in the case of natural languages, whose systematic regimentation does not precede, but follows, their coming into existence.

perhaps something like a language-of-thought representation encoded in it, must be immediately meaningful to me. Now, suppose that that very same representation, *A*, understood as a materially realised vehicle, were to be implanted within a laptop. That is, suppose that the portions of my brain that I have labelled *A*, with the relevant firings and connections, are transplanted from my brain into a laptop. How would the laptop gain access to the content of *A*? The laptop is made in a very different way from my brain: it does not have neurons, nor synapses. Hence, some interface would be needed, that is, the laptop would need to be equipped with some device that allows to register and translate the inputs coming from the neural configuration *A* into some representation that can be suitably handled by the laptop itself. Now, this leads us to the previous dilemma: either we rule out the laptop as a potential candidate for cognition because it is not capable of directly accessing the nonderived content of a representation, or we accept that that very same representation, *A*, has nonderived content for me, and only derived content for the laptop, so that the notion of nonderived content is relative to the system implementing the representation having it[9].

I do not see anything wrong with making the notion of nonderived content a relative, context-sensitive one, while I think that accepting the other alternative would be rather debatable. However, it would not be surprising, perhaps, if AA accepted such alternative envisaging nonderived intentionality as something absolute, not relative. After all, another important component of their proposal for a MOC is that cognitive processes should be individuated on the basis of their underlying processes (2001, pp. 51-52). As they remark, scientific enterprise aims at carving nature at its joints, discerning superficially similar phenomena on the basis of their lower-level differences (to use a textbook example: jadeite and nephrite are different minerals, although from a higher-level perspective they are both jade), and unifying apparently different phenomena in virtue of lower-level similarities (diamonds and coal are just made out of carbon). Thus, if we take the representations occurring within human cognisers to have nonderived content, then computers would not be cognisers, not only because they cannot access the content of those very representations in an immediate way, but also because their structure, both at a macro- and a micro-level, would be too different from the human one.

[9] Of course, I do not intend to suggest here that laptops are cognisers. My point is that the fact that laptops are not cognisers is not due to their inability to gain immediate access to the content of some representations with nonderived content.

There is a problem, though. Consider the previous example of my representation *A* with nonderived content, and suppose this time that *A* is transferred from my brain to the brain of another human, rather than to a computer. It may be the case that *A* would lose its content, if some form of externalism is true, for the circumstances making the case that *A* has the content it has may be altered. But let us grant for the sake of the argument that *A*'s content survives. Would that human be able to immediately access that content, so that *A* would preserve its status of representation with nonderived content in the proposed sense? Maybe yes, but only after it is appropriately connected to the rest of the other person's brain, and, since this is a highly speculative thought experiment, there is no empirical basis to predict whether the result would be the desired one. And what about animals? The way a monkey's brain is made is not too different from the way a human brain is (or at least, it is not as different from it as the CPU and motherboard of a computer are), so would it be possible to transplant parts of a person's brain, corresponding to representations with nonderived content, into a monkey's one?

Of course, these speculations are sci-fi scenarios, so they are nothing more than intuition pumps. Nonetheless, the intuition they are meant to elicit is that, if one does not want to relativize the notion of nonderived content, one should also espouse AA's claim about micro-composition. However, this seems problematic, since its most plausible reading is in terms of microphysical, basic causal properties, rather than in terms of functional structures. This is because choosing any particular level of description would bring the element of arbitrariness and intentional-agent-dependency back, in that one's explanatory interests would be pivotal. But then, one might ask: why should we take the details of the human constitution to set the bar for how a cognitive system needs to be? It seems that relying on too fine-grained, species-specific, or even individual-specific causal properties to identify the representations taking part in cognitive processes would be too exclusive as an approach, unless the previous intuition pumps all have a positive answer (so that some human individual's representations with nonderived content can be transplanted into other putative cognisers such as animals or other humans, without significant causal adjustments, and retaining the non-derivative status of the associated content all along). If other systems, such as animals or humans other than the one under examination, are to be potentially allowed as cognisers, then we need to accept that a

token of a representation can have nonderived content for a cogniser, but only derived content for some other cogniser.

2.2 Nonderived content as directly accessible content

One criticism that has been moved against AA's suggestion that nonderived content is a necessary condition for cognition comes from the opposite side of the debate over the Extended view on cognition. Clark (2005) has raised the issue of how to deal with mental representations of entities which in turn possess only derived content: do mental images of graphs, or the words thought in inner speech, possess derived or nonderived content? Words and things such as graphs are the kind of things that paradigmatically possesses just derived content, as they mean what they mean only because some prior intentionality has made this the case, say by convention or by established social practice. However, human mental life is taken as paradigmatically cognitive, and at least conscious thoughts have to meet AA's condition on nonderived content, in that they undoubtedly are cognitive states. So, how are we to reconcile these two facts, namely that conscious thoughts need to have nonderived content, but they sometimes appear to have the form of entities with derived content, as it were, for instance when one thinks in English?

Adams and Aizawa (2010a) have addressed this issue by remarking that, while it is indeed true that thoughts can, so to say, assume the semblance of entities with derived content, it is not the case that thoughts are themselves such entities. When one has in mind the image of a Euler-Venn diagram, what is taking part in one's mental life is not a Euler-Venn diagram, but the *thought of* a Euler-Venn diagram. What matters is not that the content of what some thought is about has been socially construed, but the sort of relation the thought bears with its content: 'the derived/non-derived distinction […] concerns the conditions in virtue of which an object bears a particular content' (Adams and Aizawa 2010b, p.582). In other words, what matters is that the relation that a thought about a Euler-Venn diagram entertains with its content (the relevant diagram) is of the sort envisioned by a naturalised semantics, or, to put it differently, that it is determined exclusively in other ways than by a meaning-conferring process carried out by some intentional subject, whatever that process may be. Thoughts can be about entities with non-original intentionality, without this undermining the nonderived status of the

thoughts' contents. Hence, I think that, as it stands, Clark's objection to AA's claim that nonderived content is a necessary component of cognition does not go through.

Nonetheless, what needs to be clarified is *why* AA uphold their view. What is the role that nonderived content is supposed to play, what is the contribution that entities with only derived content cannot provide that is, instead, necessary for the occurrence of cognition? A similar question has been raised by Menary (2006, 2010). Consider a verbal utterance of the word "dog", as opposed to one's thinking the word "dog" in inner speech, or the drawing of some graph as opposed to one's mental image of the same graph. Why would, in both examples, the latter terms of comparison but not the former be relevant for cognition? Simply pointing at the latter's having nonderived content in the sense of meeting the requirements of a theory of nonderived intentionality will not help, obviously, because what is at issue is precisely the sort of contribution that nonderived content is able to offer. Similarly, given that this point has been raised in the context of a discussion over the Extended Mind view on cognition, observing that the latter is inside someone's mind, while the first is an external object, would not do, as what is at stake is whether both occurrences, or just the latter, can be legitimately considered to be part of someone's mind.

Menary has framed his analysis in terms of the contents associated with the mental (dog_m) and with the verbal (dog_v) occurrence of, say, the word "dog". There are just two alternatives: either dog_m has the same content as dog_v, or they differ in content. But it seems implausible that a difference with respect to their content can systematically occur, as shown by the fact that one can think dog_m and immediately after utter dog_v, or vice versa, without there plausibly being any change in the meaning associated with the two. Hence, it seems safe to assume that the difference in relevance for the occurrence of cognition between dog_m and dog_v is not to be understood in terms of their semantic content. What does it consist in, then?

I am not sure that AA would be quite happy with the line of reasoning that I will engage with from now on, and which I have already introduced in the previous section. Nonetheless, I think that a promising starting point, if just incidental and not further pursued by them in subsequent discussions, can be found in AA's remark that, if we were to build a 'machine designed to meet the conditions of a true theory of nonderived content, the symbols would also have meanings *for the machine*' (2005, p.665, italics added).

What AA are silent upon is what having meaning *for* something amounts to. This, I think, is the crucial, and undeservedly under-appreciated, point. A representation[10] can be meaningful regardless of its having derived or nonderived content: my dog_m is as meaningful to me as the immediately following (or preceding) dog_v. But there is an important difference between the way in which a representation with nonderived content is meaningful and the way a representation with derived content is. If a representation's being meaningful means that its content is accessible, then a representation with derived content is only meaningful to someone as a consequence of an interpretive process. On the other hand, representations with nonderived content do not seem to require any interpretation for their content to be accessible.

That representations with derived content need to be somehow interpreted in order to have access to their content seems to me indisputable (think of words, pictures, paintings, signs…). After all, if one does not even tacitly know about the existence and nature of some relation that connects a representation with derived content with its content, it would not be possible, by definition, to gain access to the content of the representation in question. Representations with derived content are such just as a result of a meaning-conferring happening (a stipulation, a socially established practice, or other analogous processes), and if the existence of such happening, as well as what it concerns (i.e. which representation is related to which content), is not known by some intentional subject, that subject would have no way to access the content of representations with derived content. For instance, if I have never seen the logo of the organisation Emergency, I cannot fathom what it stands for by simply looking at it.

The point that, instead, needs to be argued for is the claim that representations with nonderived content do not need to be interpreted to be meaningful. This is a rather tricky task, for the only representations with nonderived content I can think of are thoughts, and it is a reasonably established fact that introspection is not particularly reliable when it comes to these matters. In fact, there might be interpretive processes which are not introspectively accessible to me. However, if such interpretive processes occurred, it would be important to establish what would be interpreted, and what would perform the interpretation. In saying that thoughts have nonderived content, I am referring, as a

[10] Remember that for simplicity I am calling entities possessing intentional content "representations", although they generally do not need to be such, properly speaking.

minimum, to consciously occurring thoughts. So, if I focus on the mental image of an apple, I am not aware of being interpreting my mental image, and if any interpretation is taking place, it must happen at the sub-personal level. But if it does, then at some point the result of such interpretation has to be made available at the personal level, or otherwise I would not be consciously entertaining the mental image of an apple. However, I am of course not aware of being introspecting any sub-personally available product of such an interpretation. Hence, given that there is no interpretation at the personal level, there are two possibilities. Either the result of the interpretation occurring at the sub-personal level is immediately made available to my conscious self; or, if such result remains only available at the sub-personal level, there needs to be, in turn, some interpretative interface at work for that result to be made available at the conscious level. Clearly, neither of these alternatives offers promising prospects for the claim that representations with nonderived content need to be interpreted. For, if the latter is chosen, then we would have a regress of interpretive processes. But if, instead, the former were correct, then there would be a stage at which the content of one's own mental states is given immediately, without any interpretation. This would make the presupposition that the mental image of the apple undergoes, in order to be meaningful, a sub-personal interpretive process superfluous in the first place. The reason is that the whole point of assuming the existence of an interpretive process at the sub-personal level was to undermine the alleged immediacy of the givenness of the content of the mental image to the conscious self. In short, the possible existence of interpretive processes at the sub-personal level does not seem to affect the claim that, at least at the personal level, nonderived content does not need interpretations to be accessed.

If one were not yet persuaded by the fact that at least consciously entertained thoughts, the paradigmatic example of representations with nonderived content, do not need to be interpreted, one might reflect upon the very notion of nonderived content. So far, we have seen that AA depict nonderived content as the content possessed by a representation without the need for some prior intentionality to establish this intentional link. Now, if a representation has nonderived content, it means that the link between that representation and its content must be such that it automatically occurs, and that its establishment would obtain even without the contribution of any intentional subject other than the bearer of the representation with nonderived content. Therefore, while in the case

of representations with derived content an interpretive process is needed to make the link between the representation and its content effective, in the other case the link cannot depend upon an interpretive process, because this would require the existence of an interpreter, that is, of an intentional subject. A merely mechanical translation does not count as an interpretation, so a non-intentional subject would not suffice.

If what I have said so far is correct, we have made one step forward in understanding the notion of nonderived content. In fact, AA's account of nonderived content just as original intentionality does not shed light on why it can be taken as a necessary condition for cognition, because it is not clear how the occurrence of processes with nonderived content would affect the instantiation of cognitive phenomena. On the other hand, the present proposed understanding of nonderived content also as content which is accessible without the need for interpretive processes is more telling, if only because the understanding I am advocating subsumes the canonical one. If it is correct, then it follows that it makes a difference to instantiate nonderived rather than derived content: it means that cognitive processes need not interpret the representations they involve and that if a process needs to do so in order to gain access to their contents, then it does not qualify as cognitive.

3. Nonderived content requires the existence of a subject

It will be helpful at this point to briefly summarise this analysis of AA's necessary condition on cognition. According to them, cognitive systems must involve representations with nonderived content. This means, on the one hand, that the content of such representations must be possessed by them independently from the influence of some pre-existing intentionality. On the other hand, and relatedly, this means that for this to be possible it is necessary that there is no interpretive mediation between the representation and the cognitive system for which it has nonderived content. Finally, the same representation can have derived content for one system, but nonderived for another, and vice versa.

Now, consider Searle's (1980) famous Chinese Room thought experiment, in which Searle-In-the-Room (SIR), who does not know Chinese at all, is given cards with Chinese symbols and, following the instructions written in English in a rulebook he is equipped

with, sends outside the room other cards on which he has written the symbols that the rulebook tells him to write, based on which symbols are written on the cards he receives. One of the results of this thought experiment is that a purely formal manipulation of symbols (representations with derived content), even if performed in such a way that its outcome is not distinguishable from the behaviour that a cognitive system would display, is not sufficient for the occurrence of a mental life of the relevant sort. Although from an external, third-person perspective (3PP) what SIR does is indistinguishable from what a competent Chinese speaker would do, we know *ex hypothesi* that, from SIR's first-person perspective (1PP), SIR has a different understanding of what he is doing. In fact, if SIR is a cogniser (in the original thought experiment, he is human), then the cognitive process he is performing consists in applying a rule (we may assume that he has memorised the entire rulebook, so that he immediately knows what to write on the card) and producing outputs on the basis of determinate inputs, both in the form of uninterpreted scribbles on cards. SIR is not engaging in a conversation, or better, SIR is indistinguishable from somebody doing so from an external, third-person perspective (3PP); but this is not the case from his own first-person perspective (1PP). Why not? Because in order to do so SIR would have to be able to recognise that the scribbles are representations, and to interpret them so as to have access to their derived content by means of the formation, within the cognitive system that he is, of representations that have nonderived content.

In an alternative version of this thought experiment, Searle replaces SIR with a purely mechanical device, entirely made of pipes, levers and pullers; call it SIR_2. SIR_2 does not seem to be in a different situation from SIR: from a 3PP it is behaviourally equivalent to a cogniser engaging in conversation, but from its own 1PP it is not performing the cognitive task it is assumed to be performing. However, most people are intuitively reluctant to grant SIR_2 the status of cognitive system, while we commonly take SIR to be a cogniser. Why? In both cases, uninterpreted external inputs are converted into something that can be manipulated by the system, which, by means of a mechanical procedure, produces some outputs. One might claim that the difference is that what is going on within SIR is the processing of representations with nonderived content, while what is going on within SIR_2 is a purely causal process, where the subsequent states of the system are not even representations, let alone representations with nonderived content. But why would this be the case? After all, these states might be understood, from

a 3PP, as representations, and if that is the case, then these would have nonderived content, because if SIR_2 actually were a cogniser, then they would probably be immediately meaningful to SIR_2 as his own neurological states are meaningful to SIR[11].

It seems to me that, if the occurrence of cognition cannot be established exclusively in terms of what can be observed from a behavioural-functional point of view, i.e. from a 3PP, then it is only possible to establish it if the right sort of first-person-perspectival happenings take place. But then, if we understand the first-person perspective as something closely related to introspection, the cognitive system under examination itself is not just the only judge for its own conformity to AA's nonderived content condition, but it also needs to be the kind of thing that can play that role. And if that is the case, then it may seem that nonderived content is not all that is required for cognition: the occurrence of a first-person perspective, even in some embryonic form, is required not just for the detection, but also for the occurrence of nonderived content, and thus for cognition. The deeper problem emerging from the previous considerations on Searle's thought experiments does not only have to do with the difficulties concerning the detection of cognition, but also with the very presence of cognition. The problem is that if the right first-perspectival facts are not in place, then there can be no cognition, regardless of what an external observer can tell from a 3PP. Hence, assuming that nonderived content is necessary for cognition, and if my proposed understanding of it is reasonable, then nonderived content needs to be complemented by another necessary condition for cognition, namely the existence of a first-person perspective.

How can we make sense of this, if the 1PP is taken to be something closely related to introspection, as it seems commonly taken to be? The risk is to fall for the well-known homunculus fallacy: a cognitive system attends to its own inner states as if it were an external observer, but from the inside, so to speak. This would only lead to a *regressus ad infinitum*, in that these dynamics would need to be reiterated, because the states of the cognitive systems attending to its own states would need to be attended by the cognitive

[11] It is worth noting that a common functionalist objection to Searle's thought experiment is that neither SIR_1 nor SIR_2 should be taken to be the relevant candidate cognisers here: the room as a whole should be instead. While this is a reasonable reply to the original thought experiment, I do not think that for the present purposes it makes any difference. Even if the room as a whole were the candidate cognitive system, then one could raise again the question of why it seems counterintuitive to attribute cognitive status to the room as a whole, while it is not counterintuitive to grant cognitive status to a functionally equivalent human being. Such question would lead to the same considerations as the ones relative to SIR_1's and SIR_2's case.

system, and so on. I think that a source of inspiration can be found in phenomenology. One elementary component of the phenomenologically conceived conscious experience, which I will refer to as "self-referentiality", can be synthetically characterised, as reported by Zahavi and Parnas (1998), as the basic modality in which the 1PP is accessed by the cognitive subject, that is, as the mental life of the subject *qua* mental life of that subject, and not of someone else or, more generally, as something external. In a nutshell, conscious states are by their nature given as belonging to whoever enters in such states.

Of course, phenomenology is mostly concerned with consciousness rather than with cognitive phenomena more broadly construed. However, it seems to me that the core idea underlying the concept of self-referentiality can be usefully applied in the context of general cognition. Accordingly, in what follows I will suggest that the possibility to access the nonderived content of a system's own representations in an immediate way can be understood as being logically, although not ontologically (this would be committing the homunculus fallacy), preceded by the very existence of a subject, i.e. of an entity capable of entertaining a 1PP, to whom those representations belong.

If the representations of a cognitive system cannot be referred to the system in question, their content cannot be accessed in the way nonderived content can, but it would be given as something external, accessible by means of some interpretive process. To see this, consider pathological cases, such as some severe symptoms manifested by people affected by Dissociative Identity Disorder (DID). People affected by DID have most frequently experienced some major trauma (such as physical or sexual abuse, perceived as overwhelming and without any possibility to escape, during their childhood), and in order to cope with it, their self has been somehow split into different, generally completely distinct personalities[12]. These cases are especially interesting for the topic at hand, in that the same physical medium (the same brain) is the realiser of a plurality of largely non-overlapping cognitive subjects, each of which has cognitive states that meet the nonderived content condition, but only with respect to the specific subject having them. In fact, the cognitive states of the various alter egos appear to belong to the single personalities in pretty much the same way in which the cognitive states of a person belong to that person alone, and not to other people. The cognitive states of one personality are

[12] According to Ross and Ness (2010), this phenomenon is not to be understood as 'an anomaly or an aberration' (p. 465); rather, on the basis of their study, it has to be considered as an extreme manifestation of the normal human pattern of response to stress.

given as belonging to that specific personality, who is seemingly capable of accessing their content in an immediate way, thus meeting the nonderived content condition. Correspondingly, those states are not given as belonging to other personalities, who are accordingly not capable of having a proper 1PP on them, nor to access their content as nonderived[13].

I think that reflection upon DID, or on some of the most severe symptoms of schizophrenia where patients report to be attending to their own thoughts and bodily movements as if they belonged or were performed by someone else (as reported by Zahavi and Parnas (1998)), hints to the hypothesis that a pre-condition for the nonderived content condition itself to be met is that the representations involved in cognitive processes display the primitive, pre-reflexive, passive (in the sense of involuntary and automatic) self-referentiality which phenomenologists have been concerned with. In other words, it seems to me that a more basic necessary condition for meeting AA's necessary condition on cognition is that cognitive representations are given to the cogniser as belonging to the cogniser itself. To be clear, this does not amount to the claim that the pre-existence of a subject is needed for cognitive states to occur. Rather, the relation between a subject and their states with nonderived content ought to be conceived of as one of simultaneous complementarity. On the one hand, nonderived content needs to be referred to a cognitive subject in a first-person perspectival way; on the other, for something to be a cognitive subject entertaining a first-person perspective, that something needs to enter into states endowed with nonderived content. Nonderived content and the first-person perspective seem unable to occur independently from one another, and they both are required for the occurrence of cognition.

I will return to this later. For now, suffice it to point at the fact that the requirement on nonderived content appears to need to be supported and integrated by the notion of first-person-perspectival subjectivity.

[13] It is however important to report Dell's (2006) observation that the characterisation of DID I have just offered only corresponds to one of two clusters of dissociative phenomena. It is in fact much more frequent to encounter cases of DID manifesting dissociative phenomena pertaining to the second cluster, i.e. cases where rather than a complete switch of personalities there are intrusions in the mental life of the main personality by other personalities.

4. Adams and Garrison and the system's reasons

I think that the aforementioned point concerning the necessity for cognition of the existence of a subject which the representations having nonderived content belong to can be applied also to another plausible necessary condition for the occurrence of cognition that Adams and Garrison (2013) have recently suggested. According to them, cognitive processes can only occur in virtue of the system's own reasons. For instance, let us grant that a hedgehog's defensive behaviour can be considered an example of cognitively driven behaviour. Although there might be an evolutionary explanation for hedgehogs' rolling into a ball when confronted with potential threats, this explanation can at best address the predisposition of hedgehogs to engage in such behaviour, while the explanation for some individual hedgehog's rolling is to be given in terms of the internal states of the system. If a hedgehog does not sense the presence of a threat, it does not roll into a ball, even if that would be, from an external perspective, the appropriate behaviour under the circumstances it is in, in the light of the evolutionarily established predisposition to do so.

In short, Adams and Garrison (AG) argue that cognitive processes can only occur for reasons determined by the particular internal states of the system instantiating them, and not for reasons external to it such as species-level reasons, or for reasons had by a potential manufacturer of the putative cognitive system at issue. It will thus be appropriate to elaborate in some more detail on how exactly we should understand the notion of "reason", which is treated essentially as a primitive one by AG.[14]

At a first approximation, something may be the reason for the occurrence of some event if it is what causally led to the instantiation of such event. However, this way of understanding reasons as nothing more than causes does not seem to account for what we take reasons to be in the context of cognitive phenomena. If the glass of water I am holding right now slipped from my grip, and as a result of hitting the ground it shatters, both its slipping and its impact with the ground would cause the glass to break. Nonetheless, they are intuitively not the reasons why that happened, at least not in the same sense as a specific intellectual interest would be the reason why a certain question

[14] Of course, what follows is not intended to be an exhaustive account of what reasons are. Rather, it is just meant to be a plausible way to understand the notion of "reason" as it appears in AG's proposal.

is asked. Purely causal and physical processes do not qualify as reason-driven processes in the sense relevant for present purposes. Something more is needed.

It seems to me that there are, broadly speaking, two sorts of reasons relevant for present purposes: reasons consisting in already occurring states of affairs, and reasons consisting in states of affairs whose occurrence is in some sense (in)desirable. In some cases, an event happens for a reason if a causal process involving that event is started as a consequence of the occurrence of some state of affairs. For instance, the reason why I hang up in the middle of a phone call is because I got mad at the person I am talking with. This is an example of the first type of reasons. In some other cases, an event occurs for a reason if a causal process involving that event is started in order to make something else (not) happen. For instance, the reason why I insert a coin inside a vending machine is that I want the machine to drop a snack. This is an example of the second type of reasons.

The common feature that these two types of reason share is that the subject of the reason in question (me, in both examples) engages in some behaviour because certain states of affairs are presented as intimately related to the subject, either as presently occurring or as meant to occur as a desirable consequence of the relevant behaviour. More specifically, the reason-holder needs to somehow be aware of the fact that his or her reasons are *his* or *hers*. If I do not know, even implicitly, that *my* inserting the coins in the vending machine can bring about the outcome desired by *me*, my inserting the coins in the vending machine would not have *my* desire to have a snack as a reason. Similarly, if my irritation does not appear immediately to be mine, that is, if I am not feeling my anger without forming any (higher-order) conscious belief about this, my hanging up the phone would not have my anger as a reason.

It should be clear, at this point, what I am driving at. My suggestion is that, as in the case of AA's necessary condition on nonderived content, also AG's necessary condition on internal reasons can only be met if some self-referentiality is embedded in the putatively cognitive behaviours adopted by the system under examination. For example, consider the two following cases: a dog growling at the owner who is trying to rescue their half-chewed shoe, and a phone whose screen unlocks when its owner puts their finger on the fingerprint reader. The difference between these two processes, according to my suggestion, consists in the fact that the dog's behaviour can be explained in terms of reasons related to the dog's desires, instincts and so on. In the second case,

instead, the behaviour of the phone cannot be given a reason-explanation in the relevant sense, because the tool itself has no reasons: its behaviour is just purely causally determined, in such a way that no subject(ivity) is involved.

To be sure, one might object that the phone's internal states are, after all, the phone's own ones, and if, in some sense, they were not given to the phone as such, the causal process triggered by its owner's fingerprint would not have occurred, for instance, if the signals sent by the fingerprint reader had not been interpreted by the CPU as such. This is an important point, and in the next section, while examining Rowlands's (2009) set of jointly sufficient conditions for cognition, I will be directly concerned with this issue.

5. Rowlands' account of the Mark of the Cognitive and a preliminary proposal

While AA's and AG's conditions on cognition were supposedly necessary ones, Rowlands's (2009, 2010) set of four conditions is only meant to be sufficient for the occurrence of cognition. In this section, I will in particular relate his fourth condition to the remarks from the previous sections on subjectivity.

Let us begin by outlining the four conditions that grant cognitive status to processes meeting them (Rowlands (2010, pp.110-111, italics in the original)):

> A process *P* is a *cognitive* process if:
>
> (1) P involves *information processing* – the manipulation and transformation of information-bearing structures.
>
> (2) This information processing has the *proper function* of *making available* either to the subject or to subsequent processing operations information that was, prior to this processing, unavailable.
>
> (3) This information is made available by way of the production, in the subject of P, of a *representational* state.
>
> (4) *P* is a process that *belongs* to the *subject* of that *representational state*.

The notion of information processing involved in the first condition is a technical one that has to do with the reduction of uncertainty concerning what states of affairs are the case. However, since it is not important to dwell on the exact nature of information processing for present purposes (I will do so in the next chapter), we can provisionally, albeit

improperly, understand the term "information" in a commonsensical way. Similarly, whether the information processing involved in cognition has the *proper* function[15] of making available either to the subject or to the subsequent operations the relevant bits of information, as opposed to more generically having the function of doing so, is not an important matter to settle here.

The third condition is closely related to the first, and, just like the first, it is unproblematic, unless one wishes to maintain that cognition may not involve representations. Of course, if one were to maintain such a view, then one would probably take issue with the first condition as well, but that is not a dispute I want to take sides on here. In fact, especially in the light of Rowlands's overall conception of cognition as a "disclosing" activity, I think that not much hangs on whether such disclosure is spelt out in representational or nonrepresentational terms. The gist of the third condition is just that cognitive processes establish a bridge between a cogniser and the world, and that such bridge consists in the cogniser's somehow being affected by the world. This much seems more than reasonable to me.

The fourth condition raises a point closely related to the issue of subjectivity I have been discussing so far, in that it states that cognitive processes belong to the subject of the representations which, based on the previous three conditions, make the information involved in cognitive processes available to the cognitive subject (or to subsequent processing). The crucial bit of this condition has to do with the word "belong": what does it mean for a cognitive process to belong to a subject? Rowlands (2010) is well aware of the difficulty of finding an answer to this question, and nearly half of his book is devoted to the issue. One possible reply is that the subject of a cognitive process is defined in terms of the subject of representations with nonderived content, as it is the beneficiary of the content made available by such representations. It follows, then, that for a cognitive process to belong to a subject is to be the process that makes some information available to the subject (or better, in Rowlands' terms, that has the proper function of doing so). This is, very synthetically put, what Rowlands maintains. The core issue then has to do with how to identify a subject.

[15] The notion of proper function is an etiological one (Millikan (1984, 1989)). Some function *f* is the proper function of some entity *e*, as opposed to an improper function, just in case *f* is *e*'s function either as the result of an evolutionary process, or because this is the case by design. In other words, a proper function is an *intended* function, in either a literal or figurative sense, that is, the function that something is meant to perform, as opposed to happening to perform and to actually succeeding in performing.

Drawing on the points raised in the previous sections of this chapter, we might be able to elaborate a preliminary criterion for establishing the boundaries of the subject of cognition, both from a 1PP (first-person perspective) and from a 3PP (third-person perspective). The guiding principle is the intuitively agreeable one that a subject of cognition, both from a 1PP and a 3PP, has to be unified, and that the subject of cognition from a 1PP has to coincide with the cognitive system as characterised from a 3PP. Crudely put, we need a criterion that accounts for the fact that my brain (or body, or extended system), which is separated from that of anybody else, is associated with my mind, and not with the mind of someone else. That is, the criterion that we are looking for is to be one that simultaneously draws the boundaries of a cognitive system from a 1PP and 3PP, in such a way that, in some sense, these two perspectives are not in conflict.

My suggestion is that a cognitive system can be identified in terms of the continuous processing of nonderived content. Of course, this suggestion is very close to AA's claim, but it differs from it in virtue of the specific understanding of nonderived content that I suggested earlier in this chapter. Let us briefly recall this notion of nonderived content. The content of a representation is nonderived if it has not been attributed conventionally to that representation, *and* it can be accessed in an immediate way by some cogniser, but only in a mediate way, by means of some interpretive process, by other cognisers. Now, recall the parallelism that I have earlier drawn between the self-referential character of the 1PP and the way in which nonderived content is accessed by the subject of the relevant representations. Based on that parallelism, if the content of some information-bearing structure is nonderived with respect to some cogniser, then that cogniser, from a 1PP, is identified by the totality of the representations that can be accessed in such a way that they can be said to have nonderived content. Only nonderived content is made immediately available to a subject, and, therefore, only those representations possessing nonderived content can belong to the relevant subject. Furthermore, if some representation belongs to a subject, then that representation has nonderived content. Therefore, the boundaries of a cognitive subject, from the perspective of the subject itself, are determined by the presence of nonderived content: where the processing of nonderived content stops, the boundaries of a cognitive subject are drawn.

This, however, is a first-personal criterion, as whether the content of some representation is nonderived or not partially depends on whether it is available to the

subject in question, from the subject's perspective. A third-personal, parallel criterion should thus be offered: the boundaries of a cognitive system, from a 3PP, are set by the material "interfaces" that allow the information being processed to be transmitted via representations with nonderived content. For instance, the way my cognitive states, which have nonderived content for me, are materially realised, is not such that my laptop can immediately access their content. In some sci-fi scenario an interface would still be needed, because my laptop would need some specific device to register the inputs coming from me, and then to translate them into representations of the appropriate sort. Conversely, I cannot access the contents of the representations within my laptop in an immediate way. I need them to be made accessible through some other representational form; this is indeed the purpose of having them translated into images on a screen (the interface), as this allows the formation of a (visual) representation whose content is nonderived with respect to me.

This way of demarcating the boundaries of a cognitive system has the virtue of linking the 1PP with the 3PP, by means of the hypothesis that the notion of nonderived content can be associated with the absence of interruptions in the information-processing activity. This seems to account for the internalist intuition, expressed in the context of the criticism of the Extended Mind view, that a mathematician using a calculator has only recruited the calculator as an external aid to cognition, and that there is not an overall "Mathematician + calculator" cognitive system. Moreover, this demarcation criterion does not only play the negative role of excluding things from cognitive systems, but it also seems to accommodate the positive existence of cognitive subjects. For instance, I am an individual cogniser because all my cognitive states are given as mine, they have their self-referential character only for me, and if they had it for someone else as well, that someone would actually be *me*. This criterion, in fact, does not preclude the possibility that extended cognitive systems exist. All that is required to generate an extended system is that the representations involved in the relevant cognitive processes have nonderived content for the entirety of the candidate extended system. This is because a single subject cannot be the subject of representations with partly derived, and partly nonderived content, in the light of the characterisation of the nonderived/derived content distinction that I have given. Derived content is content that needs to be made accessible in an immediate, hence nonderived, guise, before being able to take part in the cognitive

subject's states. Therefore, cases of extended cognition are possible, as long as there is no interruption in the continuity of the processing of nonderived content between the components of the extended system.

One might object that in some cases there is a mismatch between what the cogniser believes to be part of itself from its 1PP, and what we would be inclined to take as a constituent of the cognitive system from an external, third-personal perspective. This can happen in two ways: either the boundaries of the cogniser are more inclusive from a 1PP than what it seems from a 3PP, or the 3PP on the cogniser delivers a more inclusive demarcation of the boundaries of the cognitive system. Examples of the first case are represented by the phantom limb syndrome, where someone still painfully perceives the presence of a recently removed limb, or by cases in which suitably trained people can start perceiving the tools they use to tap their surroundings as proper extensions of their body. In both examples, the cogniser's 1PP on the boundaries of their cognitive system is wider than what observers from a 3PP would imagine.

In most of these cases, we should, perhaps, give priority to the 3PP. The case of the phantom limb syndrome can be explained away as a misrepresentation, and this would not be particularly surprising, in that a subject can easily be mistaken about their own mental life, as critics of introspectionism and autophenomenology, such as Dennett (1991, 2001, 2007), have argued. The case of the stick, which is not felt as an external tool, but almost as if it were part of the body, can be explained in an analogous way by the fact that the subject wrongly attributes the nonderived content of her representations to something where the same piece of information is contained as derived content. The phenomenon in question can only occur after a long and systematic training, and, until its insurgence, the information transmitted from the stick to the hand is accessed only in a derived way: the subject simply is not aware of the interpretive process anymore, just like when one simply reads the words on the page of a book, without being aware of the interpretive process that is occurring in order to access their meaning.

What about the second sort of cases, where the 3PP offers a wider outline of the cognitive subject than the cogniser has of itself, in the sense that at least some of the patterns of neural activation that are relevant from a 3PP do not correspond to states which the cogniser can have access to? Consider people affected by DID. In those cases, there allegedly is a plurality of cognitive subjects where there normally would only be one. It

is important to stress that the various personalities that can co-exist within the same body are not necessarily aware of the existence of others, and they generally interact with each other (if at all) in pretty much the same way different individuals would. The thoughts of one personality are not the thoughts of the others, and the thoughts of the others lack the self-referential character that the thoughts of one personality have for that personality itself. And yet, the brain hosting all of them is just one. How can we make sense of this in terms of the criterion that I have suggested? It should be noted that a process that could give rise to cognition, but which actually does not, should not be considered a proper cognitive process, but only a potential one. Hence, although a thorough neurological explanation for DID is not yet available, my guess is that, just like in cases where people do not recognise their ownership of some part of their own body, something must have gone wrong, so that some sort of "barrier" disrupting the continuity of the processing of information has come into existence.

This last remark raises an important question: how are interruptions of information processing to be exactly understood, from a 3PP? It can be rightfully pointed out that the processes going on in a thinking brain are a motley crew. There is not a single way in which neurons interact with each other: why does this not lead to an interruption of the "nonderived content chain"? This question, especially if thought in relation to the DID example, concerns precisely the extent to which some variation from a standard causal transmission of representations can occur without this leading to the coming into being of an interface. For this reason, it is not helpful to appeal to some abstract criterion applying homogeneously across the board. The best course of action, at least at this point of the inquiry, is to wait and see what empirically informed research will have to say about this.

Summary

In the first section of this chapter I have been concerned with the first of the three representative extant MOC's I have taken into consideration, namely Adams' and Aizawa's claim that it is necessary for cognitive processes that they involve representations whose content is nonderived. I have started by presenting their view as they report it themselves, and I have then examined some objections moved to their

account (specifically, Dennett's one, and in the following section also Menary's and Clark's ones).

In the second section, I have started my own analysis of this first attempt at providing a necessary condition for cognition. In particular, I have raised two points. The first one concerns the very notion of nonderived content, which, I have argued, appears to be not an absolute notion, but rather something which has to be considered with respect to some cognitive system. That is to say, I have claimed that a representation having nonderived content for a given cogniser may only have derived content for some other cogniser, and vice versa. My second point was instead about a shortcoming of AA's account concerning not its feasibility, but the motivations which may lead one to accept the processing of representations with nonderived content as a necessary condition for the occurrence of cognition. If one is to understand the notion of nonderived content solely as original intentionality, i.e. a form of intentional content which is not possessed as the result of a more or less arbitrary attribution by an intentional subject, then it is hard to see what its contribution towards cognition would be, so as to justify its supposed necessity. My opinion is that we ought to understand nonderived content not only as original intentionality, but also as content which is accessed by the bearer of the related representations without having to interpret them.

In the third section, drawing upon the conclusion of the previous section, I have suggested that nonderived content, if it is necessary for cognition, and if it is so because it is immediately accessible content, requires the existence of a subject. The reason for this is that, if there were not a sort of self-referentiality which allows the access to nonderived content to be not only immediate, but also intrinsically given as belonging to the cognitive subject, then we would risk falling into the homunculus fallacy. Furthermore, the psychiatric disorder currently known as DID appears to corroborate this view: in order to be immediately accessed, nonderived content needs to be given as belonging to the relevant cognitive subject. Representations having nonderived content for a subject would cease to have it if, for some reason, they were given as non-self-referential, where the self in question refers to the aforementioned subject.

In the fourth section, I have briefly commented upon a second proposal on the mark of the cognitive, namely Adams' and Garrison's view that another necessary condition for the occurrence of cognition is that cognitive processes be performed for the cogniser's

own reasons. My main point about their proposal was that, just like AA's one, it appears to require a notion of subject understood as something to which the reasons can belong.

Finally, in the fifth section, I have presented Rowlands's MOC. Starting from focusing my attention on the fourth condition, which explicitly mentions the notion of a subject of the representational states possessing nonderived content, I have used Rowlands's MOC as a starting point for some further reflections on the relation between subjectivity and cognition. In particular, I have argued that a plausible way to understand what it means for a cognitive process to belong to a subject is in terms of the continuity of processing of representations with nonderived content.

Chapter 2

Introduction

In this chapter, I will argue in favour of the claim that cognitive phenomena need to involve a first-person perspective (1PP). This is a claim that I have already discussed in the previous chapter; hence, the purpose of this chapter is to follow a different route to the same claim. I will start from some methodological considerations, showing how one of the difficulties associated with formulating an account of cognition consists in drawing a sharp line in the continuous evolution of biological systems, separating those organisms whose constitution is sufficient for the occurrence of cognition from those whose constitution is not sufficient. Then, in the second section, I will consider one minimal, necessary condition for cognition, namely that, as also proposed by Rowlands, information processing is a core feature of cognition. Drawing on the reflections of the first section, I will examine whether it is possible to give up on the intention of drawing a line separating systems possessing merely necessary features to earn them cognitive status from systems possessing also sufficient features. That is to say, I will explore the possibility that the minimal necessary condition for cognition is also a sufficient one.

My assessment will be a negative one: even if there is a sense in which accepting information processing as not only necessary, but also sufficient for cognition, may be less unwarranted a move as it may seem at first, it remains a problematic choice nonetheless, as it would make cognition inherently indistinguishable from all sorts of other natural phenomena, undermining its status as a natural phenomenon in its own right. Thus, I will conclude this chapter by putting forward a more specific claim about what a more satisfactory account of cognition should look like. Namely, I will suggest that cognitive phenomena may need to involve a first-person perspective (1PP).

1. The heap of cognition

Our understanding of cognition is vague, since we currently lack a generally agreed-upon set of necessary and sufficient conditions for something to qualify as cognitive.

However, providing one seems to necessarily require us to draw a line somewhere, so that some entities in the world make the cut, whilst others do not. The difficulty consists in settling for a criterion which accommodates pre-theoretical intuitions pulling in different directions, without being arbitrary and while being scientifically (empirically) informed and acceptable. Indeed, the reason one might be interested in spelling out a mark of the cognitive is not a terminological one. The purpose of finding a set of necessary and sufficient conditions is not merely to establish how we should use the word "cognition", but to capture the nature of a phenomenon which we take to actually occur in the natural world. As a matter of fact, we take some things to be paradigmatically cognitive, some other things as dubious cases, and others again as not cognitive at all: we want to know what makes this the case.

A problem that one immediately has to face is whether this issue has to be approached from a higher- or lower-level of description. That is, in analysing undisputed examples of cognition such as human cognition, should we focus on providing a set of criteria framed in functional terms, or should we focus on the description of the exact neuro-physiological phenomena underlying cognition in paradigmatically cognitive systems? Here, I will start approaching the issue from a broadly low level of description, reflecting upon cognition as an evolutionarily determined phenomenon[16], and I will end up taking the discussion to a higher level.

Organisms have observable features, called phenotypes. The evolution of phenotypic traits is a very slow process which typically occurs across numerous generations. This is because phenotypic traits result from the complex interplay between genetically determined causal predispositions and environmental influences. The genetic makeup of a given organism is, very broadly speaking, a combination of the genomes of its parents. However, internal processes, as well as influences external to the organism itself, lead to the recombination or the mutation of such genetic makeup, so that an organism's genetic makeup, although stemming from the union of its parents', is not just a straightforward sum of half of the mother's DNA and half of the father's. In short, the offspring's genes are a combination of their parents' ones, but they also differ from the

[16] For the following sketchy genetic and evolutionary considerations I am drawing from the first three chapters of Buller (2006).

received genetic material to some extent. Such differences might or might not lead to observable differences in phenotypes.

Some genetic mutations result in adaptive traits (phenotypic traits which make an organism fit to survive in its particular environment), some result in maladaptive traits (traits which make the organism less fit for survival in its particular environment), and the vast majority is neutral with respect to adaptiveness. However, it is never the case that a single genetic mutation is able to lead, on its own, to an actual adaptation, be it adaptive, neutral, or maladaptive. Even though the presence or absence of particular genetic letters in specific loci might be crucial for the manifestation of some phenotypic trait (e.g. the presence of a single chromosome X or Y in a specific locus is crucial for human beings to be, respectively, female or male), the occurrence of some phenotypic trait depends on many other genetic as well as environmental contributions. There is no such thing as a gene uniquely responsible (i.e. individually sufficient) for a trait, although the presence or absence of some specific gene might be necessary for that trait's manifestation. Of course, some genes can be especially relevant for some phenotypic trait, but expressions such as "the gene for *x*", which are common and useful shorthands, should not be taken in their literal sense.

With that being said, familiar processes of natural selection lead to an organism's genetic makeup transmission to the next generation, or not. If an organism, due to accidental causes or because of displaying maladaptive traits, does not succeed in reproducing, the mutations with respect to the previous generation occurring in its genetic makeup go lost. As a consequence, only small variations can occur across generations, and a species' genome varies at an extremely slow pace. Accordingly, even phenotypic variations occur at an extremely slow pace as a general rule, with some environmentally determined exceptions which we should not be concerned with here[17].

Now, cognition is a phenotypic trait. Furthermore, it does not seem to be a peculiarly human one, as at least some other animals display it, although the specific ways in which it is manifested vary with the species under examination (pretty much in the same way as limbs are phenotypic traits, but different animals have different sorts, and numbers, of limbs). It is then reasonable to infer that the genetic makeup responsible for

[17] For instance, if a child is not exposed to language during some crucial developmental phase, it will not be able to learn to use language later on, and if an animal's eye is kept shut during its infancy, that animal will not be able to see from that eye later on, despite the eye's being healthy.

the development of structures allowing the occurrence of cognition has developed gradually, since small genetic mutations occur from one generation to another, and considering that not only humans are capable of cognition, but also at least some other animal species are. The relevant genetic makeup is not inherently specific to humans, nor to any other animal species, be they in their ancestral or present-day forms: it must have made its appearance over the course of evolution, and it must have changed over time, as suggested by the fact that there is not a single way to manifest cognition.

The upshot is not only that there is not a single genetic microstructure underlying all forms of cognition, but also that there is not a unique form cognition assumes when occurring. Consider the genealogy of an individual human cogniser. Such individual is assumed to be a cogniser, and their genetic makeup slightly differs from their parents', so that also the way in which cognition is instantiated by them is, plausibly, slightly different from the way it is instantiated by their parents[18]. However, the way some individual manifests cognition is not sufficiently different from the way their parents manifest it to disqualify the latter as cognisers. But if this is correct, then a similar line of reasoning can be applied to the parents with respect to their own parents, and so on.

It is easy to see that this approach to cognition leads to something very similar to the ancient sorites problem. Starting from one grain of sand, the addition of another grain of sand does not give you a heap of sand; but adding a third grain still does not result in a heap of sand, and so on until we reach the paradoxical conclusion that an arbitrarily large number grains of sand still does not constitute a heap. However, the direction in which the paradox develops is slightly unorthodox in the case of cognition, as the direction is not the commonly presented one. In the original paradox, the rule is "if n grains of sand do not make a heap, $n+1$ grains of sand do not make a heap", while in the case of cognition the rule seems to be "if a set of biological properties $C = \{x_1, x_2, \ldots, x_n\}$ counts as cognition, then a set of properties $C_1 = \{C - x_k\}$, where $1 \leq k \leq n$, counts as cognition too". This is due to the assumption that no single biological property x is individually necessary nor sufficient for cognition, in the sense that no individual

[18] I am neglecting environmental influences for simplicity, but the present point can be further corroborated if cognition is conceived as not exclusively depending on the physical makeup of cognitive systems. This is because if not even non-negligible changes can immediately interrupt the "sorites" chain I will mention in a moment, then blocking the slippery slope becomes all the more problematic.

biological property is uniquely responsible for the occurrence of the higher-level phenomenon (phenotype) we call cognition.

How do we proceed, then, in order to individuate a mark of the cognitive, that is, a set of jointly necessary and sufficient conditions for something to qualify as cognitive? Three observations can be made. First of all, in both the cases of the heap of sand and of cognition we have an intuitive grasp of what the target phenomena are; namely, a sufficiently high number of grains of sand, and something related to thinking, respectively. Second, in the case of the heap of sand we are aware of what the necessary conditions are (grains of sand are necessary, and there needs to be a plurality of them), and the difficulty consists in finding the precise sufficient conditions (the exact number of grains of sand). On the other hand, we have examples of sufficient sets of conditions for cognition (an exhaustive description of the properties of a paradigmatic cogniser would be a sufficient set of conditions), but we struggle to find a minimally sufficient subset of any of those sets which is also necessary. Third, differently from the case of the heap of sand, cognition as standardly conceived comes into two seemingly importantly distinguishable varieties: it can be conscious or not. This means that consciousness is not necessary for the occurrence of cognition, as there can be instances of cognition which are not conscious. For instance, much of what the human brain does is not consciously available (think of whatever neural processing underlies the conscious performance of calculations) and yet this does not disqualify such activities as cognitive.

In the next section, I will draw upon these three observations, and I will outline a characteristic of cognition which seems to be plausibly inferable from them. Namely, I will expose the claim that general cognition could be taken to amount to the causal interaction of information-bearing structures.

2. An inclusive take on cognition

The last of the three observations with which I have concluded the previous section brings us to the realisation that what is at issue in finding a mark of the cognitive is not a characterisation of cognition that incorporates conditions specifically related to consciousness. This is because consciousness is plausibly not displayed by all putative cognitive systems, and because cognition can be both conscious and not conscious.

Rather, what we seem to need to account for is primarily non-conscious cognition: it is not implausible to take conscious cognition to be understandable as non-conscious cognition plus additional constraints, so that occurrences of the former would be a subset of the occurrences of the latter. In other words, any necessary condition for cognition in general would also be necessary for conscious cognition, but not vice versa, and any sufficient condition for conscious cognition would also be sufficient for cognition in general, but not the other way around.

As I have remarked, we have an intuitive, if vague, grasp of what cognition is. Despite this fact, we take ourselves to be the prototypical examples of cognitive systems. But we are conscious cognitive systems. This is problematic, as our understanding of cognition is modelled upon too specific an example, and we do not have introspective access to our non-conscious cognising. Nonetheless, it is possible to say at least two things about our non-conscious cognising. First of all, at least some forms of non-conscious cognitive activity seem to work differently from how conscious cognitive activity works, as testified by the literature on the widely accepted theory of reconstructive memory, stemming from Bartlett's (1932) seminal work. While in entertaining some particular memory we consciously have access to the memory somehow as a whole, and we have the illusion of retrieving it without tampering with it in any way, memory really does not seem to work as a storage of unitary memories. Rather, memories simply do not exist *qua* whole, unitary memories when they're not "under the spotlight of consciousness", metaphorically speaking. They are reconstructed upon recollection, so that remembering is not merely the retrieval of some pre-existing object, but also and importantly a productive activity. This follows from two considerations. First, a memory, i.e. all the higher-level, mental states associated with the recollection of a past event, would trivially lack all of its phenomenal and, more generally, 1PP-related components if it is not entertained at the personal level, because it is hardly possible to make sense of all these features without referring them to a subject (it would be some sort of "unthought thought", which is not something we can accept). Second, this point is corroborated by the fact that memories can change depending on the particular state a subject is in when recalling them, on the knowledge the subject possesses at the time of the recollection, and depending on how distant in the past is their formation. If they were entities stably existing independently from being consciously accessed and

from the other activities going on within our cognitive system, these facts would be left unexplained.

Furthermore, there is reason to think that not only memories, but also other non-conscious cognitive states do not (always) occur unconsciously in the same way as they occur when consciously entertained. At least some standing beliefs seem to be formed in their propositional form on the spot when consciously entertained, rather than retrieved from some storage of already formed beliefs. For example, if someone were to ask me whether a flamingo is heavier than Mars, I would immediately say "no (I believe not)", without having to consciously wonder whether that is the case or not, and, importantly, without having previously done so, as I would be entertaining such a belief for the first time. A great many beliefs exist as such only when they are consciously entertained; when they are not, there only is the relevant information stored in the brain, and such information is stored in the form of neural configurations. But, properly speaking, and following a line of reasoning analogous to that I have followed for the case of memories, a standing belief is nothing more than the possibility to form the corresponding occurring belief, but it is not itself a full-fledged belief. At least, not insofar as a belief has to be propositionally expressible in a sharp way.

This leads us to the second point that can be made about non-conscious cognition. Since arguably not all non-conscious cognitive states and processes are of the same nature as their conscious counterparts, there is no need to suppose that they take place due to processes which may not be fully naturalistically accountable, in such a way that they could not be reduced to the neural goings on of the brain, as it is instead hotly debated when it comes to conscious cognitive activity. That is to say, non-conscious cognition uncontroversially takes place entirely in virtue of causal, physical events. This is because the information-bearing structures taking part in cognitive processing do not need to be attributed the mysterious nature that consciousness introspectively seems to have[19]. Neural configurations are just arrangements of physical entities, and such entities causally interact with each other in accordance with laws and mechanisms which are, at least in principle, fully naturalistically accountable for. As a consequence, if we leave

[19] I am not saying that consciousness *is* something not entirely subject to causal laws. Rather, I am just saying that from our own perspective it is difficult to conceive of it as entirely physicalistically accountable for.

consciousness aside, there does not seem to be anything more to non-conscious cognition than causal interactions of information-bearing structures.

Before moving forward it is important to briefly clarify the notion of information at play here, namely Shannon information. This is the one standardly used in information theory since Shannon's (1948) pioneering work, and it construes information as the measure of the reduction of uncertainty concerning potentially occurring states of affairs. For instance, tossing a fair coin can have only two possible outcomes, namely either heads or tails. Hence, when one registers the state of affairs which actually obtains, one reduces the amount of uncertainty concerning the result of tossing the coin, thus producing $\log_2$ (yes/no) 2 (heads/tails) = 1 bit of information, because there are 2^1 possible states of affairs whose occurrence or absence has been ascertained on the basis of such toss. If, instead, one were to randomly pick one among the four aces of a regular deck of cards, the information generated by registering that the ace of spades has been chosen, say, would be $\log_2$ (yes/no) 4 (hearts/spades/clubs/diamonds) = 2 bits, because determining that the ace of spades has been picked also entails that the three other aces have not been picked. Thus, if one picks another ace from the remaining three, that draw would produce $\log_2$ (yes/no) 3 (hearts/clubs/diamonds) = 1,585 bits of information.

An important characteristic of Shannon information is that it is, so to say, extrinsic information, as the amount and quality of information borne by any given structure depend importantly on factors other than the information-bearing structure itself. Shannon information only exists insofar as an external observer, a receiver of the transmitted information, is in place. This is because what determines the amount and quality of information carried by any information-bearing structure is the constitution of the receiver, that is, what sort of information the receiver is equipped to receive.

Consider the case of a light detector connected to, say, a radio which turns itself on when the detector can tell that the light in the room is being turned on, and which stays off otherwise. If I enter the room without turning the light on, the detector will not receive any bit of information of the sort it is equipped to react to: I am not the sort of input which is relevant for the receiver's sensor, so that, from the point of view of the receiver, upon my entering the room nothing happens, and 0 bits of information will be produced. However, under the same circumstances, I am potentially carrying a great many bits of information. For instance, if a child was inside the room waiting for someone to open a

jar for them, my appearance would carry a lot of information for them. Whatever specific features they may be sensitive to, by looking at me, they will reduce the uncertainty concerning the state of affairs they were interested in ("can I get some help to open a jar?").

Now that the relevant notion of information has been explained, and it has been characterised as an observer-relative notion, it appears that we are in a position to formulate a preliminary condition for the occurrence of cognition in general:

(1) The causal interaction of information-bearing structures is necessary for cognition.

Some remarks are in order. First of all, what reasons are there to assume that what is involved in cognitive processing has to be understood in terms of information processing? The answer is not one that can be given on aprioristic grounds. Rather, it has to do with the consideration that, as a matter of fact, all cognitive processes have to do, each in their own way, with the processing of information. Not only remembering or believing, but also perceiving, sensorimotor coordination, propositional thinking of various sorts, feeling emotions, desires, and bodily sensations (proprioceptively or interoceptively); you name it. All these kinds of cognitive phenomena appear to involve the processing of information in one way or another, if not at the level of their consciously accessible phenomenology, at least at the level of the physical structures realising them. It is only natural, then, to construe the material structures realising cognitive processing as information-bearing structures.

The second point to be addressed is that, while necessary conditions for non-conscious cognition are necessary also for conscious cognition, sufficient conditions for non-conscious cognition are not always sufficient for conscious cognition. Hence, **(1)** is meant to be a necessary and sufficient condition for non-conscious cognition, but only necessary for conscious cognition. It might turn out to be also sufficient for consciousness, but I will not be concerned with the issue here. For now, I will assume that some further condition is needed in order for **(1)** to be sufficient for conscious cognition as well, without trying to spell out exactly what that (set of) condition(s) may be.

Finally, one of the virtues of **(1)** is that it is not related to any specific sort of cogniser, in that it applies not only to human cognition[20], but also to non-human, animal cognition, and to the activities which computers perform, even nowadays. In fact, while we might take these reflections related to human cognition to straightforwardly apply to undisputed animal cognitive systems which are also conscious, one worry might be that borderline cases are unduly classified as cognitive. For instance, in the light of **(1)**, bacteria would perform cognition, and computers would as well, as both involve the causal interaction of information-bearing structures. But there are strong intuitions pulling in the opposite direction. In particular, one might want to move an objection to the effect that only entities endowed with a mind can be capable of cognition, and it seems implausible that bacteria or (present-day) computers have a mind.

3. Perhaps too inclusive?

Even if one were sympathetic with **(1)** and with its interpretation as a necessary condition, one might still want to point out that **(1)** is not all there is to cognition, so that **(1)** is not sufficient for non-conscious cognition too. The reason would not just be that it encompasses debatable cases which not everybody would count as cognitive, but also because it lets in things that seem indisputably non-cognitive. In fact, since the notion of information involved in the expression "information-bearing structure" is to be understood in a very inclusive sense (the same sense in which one says that no information can come out of a black hole), basically any object in the world can, under certain conditions, count as an information-bearing structure. And, since everything in the world interacts causally to some extent with the rest of the world, literally everything could qualify as part of some cognitive process or other. For instance, a rock rolling off a cliff into the sea could be part of a cognitive process involving, say, the cliff, the air, and the sea, besides the rock itself, insofar as they can be considered information-bearing structures relative to some receiver. Surely, we do not want this to be the case, but perhaps there are some reasons to maintain that view. As counterintuitive as it sounds, in what follows I will explain why one may potentially be inclined to argue that there really is no

[20] From here on, I will use the term "(general) cognition" to refer only to non-conscious cognition, and I will talk of consciousness when I intend to refer to conscious cognition.

reason for not wanting to allow for cognition to apply, to some extent, to virtually anything in the universe.

Let us start by pointing out that, mindlessness, just like absence of consciousness, does not constitute a good reason for attacking **(1)**, since cognition is assumed to encompass more than just conscious phenomena and the activity of creatures possessing a "mind". In order to attack **(1)** as too loose, it is necessary to provide some substantial reason to the effect that **(1)** captures more than it is meant to. Now, as I have pointed out at the beginning of this chapter, the search for a mark of the cognitive is motivated not by linguistic interests, but by the will to understand something that we take to be a scientifically acceptable feature of reality. That is to say, cognition is not just a label which we arbitrarily apply to certain phenomena, but it is meant to correspond to something really existing in the world. If cognition is to be maintained among our scientific notions, it has to refer to some non-arbitrary aspect of the world, and this aspect cannot simply amount to consciousness.

As it happens, many of the fundamental scientific magnitudes can be potentially measured with respect to any arbitrary collection of objects. In particular, I would like to take heat as a case study. Heat is a scientifically acceptable notion, and almost everything has heat. Sure, it is possible that in some portion of space there is no heat, and in that case the temperature there would be 0 Kelvin degrees. But this only shows that the notion of heat is not a vacuous one, in that it corresponds to some natural phenomenon which might or might not occur. Furthermore, the quantity of heat present in some given object or collection of objects under examination can always be measured or calculated in principle: I have a certain body temperature, my laptop has a certain temperature, and the system composed of myself together with my laptop has a certain (average) temperature.

Notably, many scientific efforts have been devoted to the search for some substance responsible for the occurrence of heat. That is to say, just like the occurrence of water depends on the occurrence of a certain kind of substance, it was in the past assumed that there was some sort of substance responsible for the occurrence of heat. Here it is not important to review the history of heat; what matters is to point out that there is no particular kind of object underlying the occurrence of heat. Throughout history, there have been lots of theories trying to account for the fact that some things feel hot while others feel cold in terms of the presence or absence of some kind of substance, but they

all were misguided. The simple reason is that heat is not, so to say, a "substantial" magnitude, one that corresponds to the amount of some substance.

Interestingly enough, looking at the history of science, one notices that it is not rare to take as a starting hypothesis that the existence of some specific sort of entity underlies the occurrence of some still not fully understood phenomenon. For instance, we currently take gravity to be one of the four fundamental forces in nature, and we know a great deal about it. However, it is not yet clear exactly *what* gravity is, and among the hypotheses under consideration, there is one according to which some still undiscovered particles, so-called "gravitons", might be responsible for the occurrence of gravity.

In short, not every physical quantity needs some specific substance to exist, although some of them do, and in general we tend to look for substances responsible for the occurrence of the relevant phenomena. Not only, but it is not the case that all apparently opposite phenomena really are different in nature. Indeed, many antithetic properties and phenomena (hot vs. cold, light vs. darkness…) have originally been taken to be distinct things, while they turned out to correspond to different values associated with the same magnitude. The reason for this sort of mistake is that the way the world appears to us, from our perspective, is not always the way we find it to ultimately be. Some things are meaningful in specific ways to us because of our physical makeup (certain light conditions are perceived as "darkness" by humans, but animals such as cats would not perceive them in the same way) or in the light of our current state (e.g. a room temperature of 20°C is perceived as warm in winter, and relatively cold in summer). But it is not always the case that phenomena which have some special relevance for us are also relevant from an objective perspective, just like it is not always the case that states of affairs which do not seem remarkable to us (the difference between how we feel water at the temperature of 1°C and how we feel water at the temperature of -1°C) are not significant under some respect (in the former case, water is still in liquid form, in the latter it is solid).

What do all these considerations lead to, in the case of cognition? My hypothesis is that, as in the case of the other magnitudes mentioned above, insofar as we do not have a firm scientific account of it, we cannot help but think of cognition on the basis of how we understand it in relation to ourselves and to what is significant for us. That is to say, I think that the claim that there is a sharp distinction between what counts as cognitive, and

what does not, relies upon the assumption that what is cognitive in general must be similar enough[21] to what cognition is in our own case. Just like it is counterintuitive to think that any arbitrary system has an overall temperature in a non-abstractly mathematical sense, it is counterintuitive that any arbitrary collection of objects might instantiate cognition in a legitimate sense. The notion of heat is such that it allows for a collection of objects individually having different temperatures to have an overall average temperature, so why could cognition not work in a similar way, on the basis of **(1)**? Cold objects seem to instantiate a very different property from the property instantiated by hot objects, while they really only have different values associated with the same magnitude. In an analogous way, it is possible to suggest that while a rock falling into the sea seems not to instantiate the same sort of property that our brain instantiates in carrying out non-conscious cognitive processes, it might actually be instantiating exactly the same sort of phenomenon (cognition as characterised in **(1)**), and that the only difference is that the two take place in different forms and degrees of complexity, which cannot be easily subsumed under the same category from our perspective of conscious beings.

There is, however, one substantial issue with this line of thought. Even granting that, in principle, any aggregation of objects can be "assessed" for cognition, so that there is no *ex ante* way to non-question-beggingly exclude certain things from the class of things that may display cognitive phenomena, and even granting that the focus **(1)** puts on information processing, we would still lack a way to discern cognition from other phenomena involving information processing. Identifying cognition with causally mediated information processing makes the very notion of cognition redundant, against the previous assumption that cognition is indeed a natural phenomenon worth being investigated in its own right. If cognition is to legitimately retain its place within our scientific theories, it does not just need to involve information processing: it needs to involve information processing that happens in specific ways, distinguishing it from other non-cognitive forms of information processing.

In short, the real problem with **(1)** is not that it may grant cognitive status to too many things, but that it could not be used to deny cognitive status to anything at all. Again, that would not be a problem if such status were more distinctively characterised, but this is not the case if we are to rely solely on **(1)**. Consider, again, the comparison drawn with

[21] This "enough" is where most of the difficulties relative to characterising cognition reside.

heat. There is nothing wrong with saying that it is possible to determine the quantity of heat associated with any region of space, but this is because heat can be characterised in a way that is unique to it. If the features that characterise heat were associated also with other phenomena (say, with macroscopic electromagnetic phenomena and with certain chemical reactions), then, whenever we have specific values associated with those features, we would not thereby just be quantifying the heat associated with the object of our measurements, but we would also be talking about the aggregate's electromagnetic properties and some chemical reactions occurring within it. This would make talk of heat, electromagnetism, and whatever chemical reactions we might have in mind, perfectly equivalent. Consequently, there would not be any need to posit the separate, robust existence of these three kinds of things as three independent sorts of natural phenomena.

To conclude, even if **(1)** may be accepted as a necessary condition for the occurrence of cognition, it is not all there is to say about cognition. It is not an individually sufficient condition. Is there, then, something more we can say about cognition that may help us elaborate an account of it, based on which cognition can be distinguished from other phenomena? In the next section, I will propose one such further condition.

4. Commonsense and functionalism

A completely satisfactory mark of the cognitive should not characterise cognition uniquely from an external, third-personal point of view. At the same time, it should not only take into consideration reports about what being a cognitive subject performing cognitive tasks feels like, or more in general seems to amount to, from the first-personal perspective of already acknowledged subjects engaging with introspection. What a MOC should instead try to do is reconciling these two perspectives, highlighting how they are related to each other.

Rowlands' MOC represents an important step towards this goal. In the previous chapter, I focused my attention mostly on the issue raised by the fourth condition of his proposal, namely the one stating that cognitive processes must belong to a subject. The reason why it is important to establish what exactly it takes to be a subject of some cognitive phenomena can be spelt out keeping in mind the importance of the 3PP on the one hand, and the crucial need not to neglect the 1PP on the other. In terms of 3PP, we

need to have a criterion which allows us to determine whether some given system is a cognitive one, and what exactly are the parts of it which are playing a role in leading to the occurrence of cognition. This is especially important if we keep in mind the modern views on cognition, where cognition is considered not to be a fully internal phenomenon, but a distributed one, which constitutively depends not just on a core which does all the work, but also on intuitively more peripheral elements. Specifically, the central nervous system is nowadays more and more frequently considered no longer to be the only part of living organisms which performs cognitive activities. Instead, there is a proliferation of views which, to different extents and in different combinations, take also other parts of the body, and even of the environment in which bodies are situated, to make central contributions towards the performance of cognitive processes.

For this reason, commonsense cannot have the last word in establishing what can be rightfully considered the subject of cognition. That we should not rely on commonsense in settling matters concerning scientific or philosophical questions will hardly be received as a bold claim. What is more debatable is the view that not even the most widely adopted approach, namely functionalism, can effectively account entirely for what a cognitive subject is and for what its boundaries are. This is because, while it is indeed a very powerful view when it comes to describing and explaining how the various parts of a given system interact with each other and contribute to the overall occurrence of cognitive phenomena, it leaves out a crucial aspect of the inquiry about what a subject of cognition is. That is to say, functionalism works very well in accounting for already given systems, but, as it emerges from the debate over the extended mind, it does not provide us with a conclusive criterion for establishing what systems should be chosen as objects of functional analyses. In other words, functionalism seems to be a conceptual tool whose purpose is predominantly explanatory, rather than heuristic: this latter aspect has to be determined in some other way. For instance, to take Clark's and Chalmers' (1998) classic example of Otto and his notebook, once one decides to give a functional analysis of the "Otto + notebook" system, one can provide more or less compelling functionalist reasons to maintain that such systems are on a par with Inga's completely internal mnemonic apparatus. But functionalism *per se* cannot give us any independent reasons for deciding to study an extended system rather than a fully internalistically conceived one. This is something that has to be done in virtue of commonsensical

assumptions (as when one assumes that Inga is a cogniser), or on the basis of some other sort of hypothesis. The reason is that, if a functionalist account of something is the explanation of what the parts of that something are in the light of the causal role they play within the overall system, then, since one can find (even very indirect or weak) causal interactions among virtually any arbitrary collection of objects in the universe, the best a functionalist approach can do is helping us to establish how relevant some parts of a system are for certain given functions, and what they do within the system. But functionalism does not prompt us to draw the line demarcating the boundaries of one system in a certain way rather than another. It can only help us in determining and assessing the causal relations among the parts of some already accepted system. To establish what the boundaries of a cognitive system are from a 3PP, and what a subject of certain cognitive processes is, we precisely need some criterion to pick some definite system and not others; only then we can engage in functional theorising about how that system works.

Furthermore, it seems that functionalism is rather inadequate if we are concerned not with the 3PP, but with the 1PP on a cognitive subject. The processes through which information is emitted, transmitted, received, and utilised in the context of cognition, considered simultaneously from both perspectives, do not seem to be plausibly understandable entirely as materialistically causal. One of the lessons that one can draw from Searle's famous Chinese room thought experiment is that the mere causal concatenations of sequences of information-bearing structures are not sufficient for encompassing all the aspects of cognition. Searle-in-the-Room[22] has no understanding of what the symbols he syntactically processes mean, and yet he manages to be functionally equivalent to a proficient Chinese speaker. From a 1PP, the knowledge of the causal interactions with which functionalism is concerned does not seem to be enough to explain what cognition is. Cognitive processes seem to involve phenomena, such as understanding, which are better understood as resulting from more mentalistically conceived, rather than materialistically construed, determination relations (e.g. that of motivation). Consequently, insofar as functionalism is concerned exclusively with the

[22] This point is meant to apply also to the functionalist rejoinder that the relevant system is not SIR, but the room as a whole. The room as a whole may be functionally equivalent to a proficient Chinese speaker, and yet it seems wrong to say that it has the same sort of semantic understanding that its human functional equivalent would have in that situation.

causal processes in which the materialistically construed information-bearing structures take part, it appears that the 1PP cannot be properly dealt with by means of a functionalist approach.

However, one may question my point concerning the inadequacy of a purely third-personal approach in explaining cognition. That is to say, one may be sceptical that the 1PP should be taken into consideration at all. After all, one may insist, if functionalism is to be considered the proper theoretical background to adopt in studying cognition, as tradition currently has it, then a machine functionally equivalent to a human (with respect to some given supposedly cognitive task, at least) should be legitimately considered to be engaged in a cognitive activity. A stronger, positive reason needs to be provided to consider the 1PP as an indispensable component of cognition. Relying on thought experiments, which are according to many mere intuition pumps, will not do. After all, trusting our intuitions concerning fictional scenarios risks being another way of passively accepting commonsense, or, more generally, some partially implicit theoretical background which needs instead to be spelt out and examined in as much detail as possible.

My reply is that, if functionalism does not address the 1PP, then it risks collapsing into behaviourism. If the 1PP were not relevant for cognition, then one would have to explain on which basis even human beings are considered to be cognisers, under normal circumstances. The reasons why we take ourselves, and other people by analogy, to be cognitive systems are, as far as I can tell, two. First, we immediately witness our own cognitive activity, so that we cannot sensibly question our cognitive status, on pain of not knowing what the target phenomenon of an account of cognition would be.

The second reason is that we have a sense of the non-incidental correlation between what happens within ourselves (i.e. from our own 1PP) and the relevant external manifestations (our third personally accessible cognitive behaviours). Had we not had a 1PP associated with our cognitive activities, and had we not had a clear sense of its connection with the 3PP, the entire issue of defining what cognition is would fade. For, not only the correlation established between certain behaviours and mental happenings would not be a spontaneous correlation, but there would not even be any mental happenings to be correlated with the external behaviours. As a consequence, entirely third-personally described cognitive manifestations would be causally driven sequences

of happenings not dissimilar to those occurring when a rock falls in a pond. There would not be anything “intelligent” about them; they would just be complex, but ordinary, natural happenings.

In short, a cognitive-looking behaviour is not really cognitive if it does not involve something that, pre-theoretically and beyond any reasonable doubt, we consider to be an ineliminable component of cognition, i.e. a 1PP. The 1PP cannot be dispensed with, in favour of an entirely third-personal one.

Chapter 3

Introduction

The purpose of this chapter is to offer more detailed, alternative arguments to the effect that, firstly, having a first-person perspective (1PP) on cognition is necessary for the very occurrence of cognition and, secondly, that having such a perspective is also sufficient for it. These topics have been already briefly touched upon at the end of the previous chapter, so the purpose of the present one is to approach them in a different way and to discuss them more extensively. In particular, I have made use so far of primarily epistemic considerations and of our pretheoretical intuitions. That is to say, I have suggested that the necessity of 1PP stems from the impossibility to account for the difference between paradigmatic cognitive systems and other systems to which we intuitively would not grant cognitive status, although they may be functionally equivalent to systems whose cognitive status cannot be denied.

Here, I will adopt a different strategy. In the first section, I will prepare the conceptual grounds for the rest of the chapter, by introducing the distinction between being a system capable of engaging in cognitive-looking activities, on the one hand, and being a subject of cognitive activities on the other. Then, in the second section, I will provide a metaphysical characterisation of exactly what the 1PP and the 3PP are, in such a way that the difference between the two perspectives will be made clear. This difference consists in the fact that the 3PP is a perspective involving a relation between two conceptually and perhaps ontologically distinct objects, while the 1PP is not relational in this way.

In the third section, I will present the actual argument for the necessity of the 1PP for cognition. My argument will be based on reflections upon cognition as a real, natural phenomenon, and I will show that adopting a purely third-personal perspective leads to a difficulty similar to the cognitive bloat objection to the Extended Cognition view. But if cognition is something which really exists, i.e. which is not a theoretical fiction, there cannot be any arbitrariness with respect to whether it occurs or not. And since the situation that would be reached by excluding the 1PP would introduce an important element of

arbitrariness in the picture, this would clash with the initial assumption that cognition is not a theoretical fiction.

1. Cognition, cognitive subjects, and the 1PP

I have concluded the previous chapter by arguing that the 1PP on cognition is just as important as the 3PP: if the 1PP were not constitutive of what cognition is, then it would not be possible to even establish what the target phenomenon is. We know that we are cognisers because we first-personally witness our cognitive activity, and we know that other humans are cognisers because we assume that they have a similar first-personal dimension. Were they to completely lack such a dimension, or better, were there not such a thing as a 1PP, then it would not only be impossible to capture what cognition is, but there also would not be any phenomenon to account for in the first place. This will be the view I will explore in this chapter, and which I will criticise and ultimately reject in the following two.

If we take the first-person perspective out of the equation, we would be left with a number of complex interactions occurring within putative cognitive systems, but there would be nothing especially puzzling about them. Consider the Extended Cognition view, for instance. If, instead of the conscious Otto, there had been a regular, present-day laptop, with a plugged-in handwriting-reader tool, it is doubtful whether any debate would have developed over whether there is a unified "computer + notebook" extended cognitive system. It would be a somewhat pedantic exercise trying to conclusively establish this, so as to be able to determine if, and to what extent, the cognitive performance that reading appears to be really occurs. The same goes for the determination of whether the computer alone is engaged in such activity, or whether we should take also the hand-writing decoder to be a constitutive part of the process, rather than just a causally coupled, but non-constitutive, external object. As another example, consider the case of the already existing self-driving cars. The issue of determining whether the "car + sensors" system is a legitimate, unified extended cognitive system would hardly be a truly problematic one. Even assuming that, for some reason, we are willing to grant that what the car does is remarkably similar to what cognisers (can) do, claims like the previous one would likely not lead to any controversy.

The reason for this might be that we do not *really* take the laptop or the car from the previous example to be cognisers. There is some pretheoretical, intuitive sense in which we as humans, but not contemporary machines, even very sophisticated ones, are legitimately considered cognisers. With this, I am not saying that this intuition is a reasonable one; what I am doing is just pointing at its presence. Plausibly, this is because we do not take being capable of cognitive-looking activities and being a cognitive subject to go hand in hand, and this seems to be the reason underlying the existence of that intuition. Curiously, being capable of engaging in cognitive-looking activities appears to be much less problematic than being the subject of such cognitive activities. As long as something engages in behaviours that seem intelligent enough, that something appears to display cognition, from an intuitive point of view. The problems arise only when we seriously take into consideration the fact that being a system displaying cognition and being a cognitive subject should not be two distinct things. The aforementioned systems, namely the "car + sensors", the "laptop + hand-writing reader", and "Otto + notebook" can be said to engage, as wholes, in *prima facie* intelligent behaviours. One may even be inclined to say that their activities are genuinely cognitive ones and that they are proper cognitive systems. But things are a bit different if subjectivity, in the sense of "being a subject of a cognitive process", is taken into consideration. That appears to be a much thicker notion than that of cognitive system. And yet, it should not be that different: a cognitive system is a collection of objects (some of) whose activities are cognitive activities, and which, as a whole, is the subject of such cognitive phenomena, that is, the owner of the related states and the author of the related processes.

What is the issue, then? The issue is that subjectivity, or, perhaps less misleadingly, the notion of cognitive subject appears to be loaded with all sorts of links to other notions such as agency, possession, privacy, wholeness and so on, which are generally associated with the more evocative concept of a mind, not with the looser one of cognition. You cannot be a cognitive subject because someone said so; you have to intrinsically be one, independently from any observer's opinion. Along similar lines, you cannot be denied the status of cognitive subject because nobody takes you to be a cognitive subject. But being a cognitive subject should not be separated from cognition. As a consequence, it should not be an observer-dependent matter whether certain systems ought to be said to engage in cognition or not. Cognition and (cognitive) subjectivity stand or fall together. So,

insofar as Otto together with his notebook is taken to be a cognitive system, in the sense that one considers the related overall activities as intelligent, it should also be taken to be, as a whole, a cognitive subject; and this clashes with some aspects of the pre-theoretical understanding of what that would mean. Analogously, a self-driving car is a system capable of carrying out typically cognitive tasks, such as navigating the world in an environment-sensitive manner, which, were they to be performed by a human (to use a formulation mimicking Clark's and Chalmers's (1998) Parity Principle), would be considered cognitive. However, it is likely that one would like to resist the implication that a self-driving car is a cognitive subject. Or, again, somnambulists may perhaps perform tasks that, were they to be performed during wake, would be cognitive. Nevertheless, it seems a stretch to say that a somnambulist is a cognitive *subject*, with respect to such cognitive performances. As a last example, consider a person with DID. They engage in cognitive activities all the time, both during dissociative episodes and during non-dissociated life, but it seems inappropriate to deem that person as a whole to be the subject of the cognitive states of any of their alters, assuming that those states are not shared among the various personalities that may coexist.

All this leads us, again, to the idea that having a 1PP appears to be an ineliminable component of cognition. To be the subject of cognition is to be the entity to which certain cognitive activities are referred; this much seems uncontroversial to me. But that is not all there is to it. Furthermore, a cognitive subject is an entity existing in its own right. It is an entity that comes into existence as soon as the relevant cognitive manifestations occur, rather than merely an observer-dependent, purely theoretical entity. Hence, there must be some non-negotiable features associated with cognitive subjectivity sustaining the existence of cognitive subjects. Moreover, these features ought to have to do with the way a cognitive subject is related to its own cognitive activities, and they are plausibly of the sort that can be best made sense if they are understood as pertaining to a first-personal dimension. For instance, a cognitive subject should have agency over its cognitive activities, and the information involved in them should be made available to it in a more intimate way than it can be accessed from the outside (from a 3PP).

Thus, I propose to introduce the following definition, where the notion of cognitive subject and that of first-person perspective are explicitly linked:

(1PP) A cognitive subject C has a 1PP on a process P iff P involves the processing of information-bearing structures whose content is made immediately available to C.

This definition *prima facie* departs from the standard way in which the 1PP is construed. Typically, saying that a subject has a 1PP on some phenomenon is understood as an epistemic claim, as it is a claim about the way in which the phenomenon in question appears to the subject: the subject has access to the phenomenon in question "from inside", rather than in the same way as an external observer would. I take the present definition to subsume this ordinary conception of a 1PP, or better, to account for the standard conception in terms of the notions I have been discussing in the past two chapters. Indeed, this definition should be understood as tying together the many remarks that I have made so far about nonderived content, cognitive subjectivity, and the 1PP. Specifically, recall my amended understanding of nonderived content. Nonderived content, I maintain, is not merely that content which is associated with certain information-bearing structures, independently from any contingent stipulation made by some pre-existing intentional agent. Rather, it is also, and crucially, that content which is made immediately available to a subject. This, in turn, means that such content has an impact on the subject's (cognitive) activities exclusively in virtue of the interaction between the cognitive system in question and the information-bearing structures bearing the relevant content, without the need for further background circumstances or information. As a result, one can make sense of what it is to have a 1PP associated with a certain phenomenon in the standard sense: it is to stand in an intimate relationship with that phenomenon, one which reflects the way the content of that phenomenon is made available to the subject.

It will perhaps be helpful to compare the following two illustrative examples. On the one hand, a written token of the word "dog", which is a representation having derived content, can only make its content available to me because I have learnt that that particular symbol refers to dogs; furthermore, I can be said to have a 1PP (in the standard sense) on my perceptual experience of the written word, but not on the written word itself. On the other hand, the neural configuration I am in when I am contemplating a mental image of a dog presumably has nonderived content, and it makes its content (the mental image of a dog) available to me (the cognitive subject) as the direct, automatic result of the

interplay between such neural configuration and the overall cognitive system (my brain). In addition to that, this seems an uncontroversial example of having a 1PP associated with a cognitive phenomenon.

These examples are purely illustrative, but the related considerations are generalisable. One can thus draw two conclusions. First, in the light of the above definition of what it is to have a 1PP, to have a 1PP on a given cognitive process can be also understood as entertaining the sort of relationship with the relevant information-bearing structures that warrants considering their content to be nonderived (in the particular sense I have discussed in the first chapter). Second, the 1PP as standardly conceived is not at odds with my definition of it; on the contrary, my definition applies at least as many cases as those covered by the standard conception.

In the next section, I will discuss some similarities and dissimilarities existing between my characterisation of the 1PP and some notions from the literature on consciousness. But, for the time being, let us return to the idea that the 1PP is an ineliminable component of cognition. The thoughts presented so far, just like the ones from the previous chapter, point in part towards an epistemically inspired reason for including the 1PP in an account of cognition. That is to say, without the 1PP, it appears to be extremely difficult to make sense of what cognition and cognitive subjects are, and to do so in such a way that a characterisation of the former nicely links it to our grasp of what the latter are. However, it seems that the argument to the effect that the 1PP must be included in an account of cognition, for otherwise we cannot properly understand what cognition is, is not the only argument for doing so, nor is it the main one. There seems to be reason to think that what is at stake is not just our conceptual grip on cognition, so that a fully third-personal characterisation of cognition would be perceived as unsatisfactory. One thing is failing to encompass an element which would allow us, given our specific circumstances of cognitive systems entertaining a 1PP on our cognitive activities, to comprehend what cognition and cognitive subjects are. Another thing is showing that, by eliminating the 1PP, not only our comprehension of cognition would crumble, but also the target phenomenon itself could not occur, and, importantly, the reason why we could not get the desired understanding would be that, in the end, there would not really be anything to understand. One can argue that the 1PP is necessary for cognition only if the

second scenario would be the one we would have to face if 1PP were left out from an account of cognition.

Thus, having the previous remarks in mind, one question needs to be addressed. How exactly would the 1PP prove necessary for cognition, in a strong enough sense that not only the intelligibility, but also the very existence of the phenomenon would hang on it? In examining this issue, the dialectic strategy I will follow will consist in working, in turn, under two opposite assumptions. Namely, I will first start by supposing that cognition is not, all in all, a genuine phenomenon; then, I will work under the contrary assumption that cognition does truly exist as a natural phenomenon.

So, let us start by assuming that we have been wrong all the time, and cognition is just a theoretical fiction to which no real, natural phenomenon corresponds. In this case, then, cognition would not be different from other theoretical phenomena which have been over time introduced in, and subsequently eliminated from, the scientific body of knowledge (e.g. Mesmer's animal magnetism). As a consequence, all we would be left with would plausibly be what is ordinarily referred to as "the mind". In this scenario, the recent developments of the cognitive sciences, where the construal of the phenomena under study has increasingly drifted away from its psychologistic (and internalist) origins, would be wrong-headed. Cognition is not something distinguishable from, perhaps even subsuming, the mind. Rather, the latter is all there is for us to study, because, if anything, minds in some sense or other do exist in the natural world, despite being mysterious in many ways. In other words, if our interest in cognition turned out to be misguided, so that whatever we found interesting and puzzling about the mind was not related to cognition, for the simple reason that cognition turned out not to exist (by assumption), then we would have to accept that all we should have focused on as the source of our interest, and as the real phenomenon at issue, was the mind in a more traditional sense. If we assume that cognition is not an actual phenomenon, attaching the label "cognitive" to certain events would be a vacuous move, on which not much depends aside from our consequently misguided theories. The legitimate object of inquiry would be something else, and if we persisted in pursuing the study of cognition as such, we could at best end up with a theory which is somewhat empirically successful, but ultimately false and devoid of any content. This would be a situation similar to the case of Newtonian physics, which is highly successful but entirely false, as the way the world works is not that envisioned by it.

So, if we assume, in the way I have just explained, that cognition does not exist, that there is not a genuine, solid, kind "cognition", then the mind is all there is, together with its correlates, be they neural or of other nature. Whatever else that is related to what we now call cognition (and which is not identical to what could be considered a mind) cannot be treated as a single natural kind in its own right. If it could, then the initial assumption under which we are working would be contradicted. Such a unitary phenomenon would simply be what we intend to refer to in talking of cognition. Therefore, in such a hypothetical scenario we should drop all talk of cognition and just focus on the mind, with all this entails. In particular, this would make the need to include the 1PP in our account of the cognitive-turned-out-to-be-the-mental all the more urgent, as it would undoubtedly be unwise, to say the least, to formulate an account of mentality which leaves out mentality as it is given to mindful beings.

To summarise, if cognition were not a real phenomenon, then the object of our inquiries would have been mentality all along, and an account of mentality must include the 1PP. No explanation of mentality which fails to acknowledge its ineliminability as a component of the target phenomenon would ever be complete, as the 1PP is a crucial, necessary aspect of mentality. Indeed, one may even go further down a dualist path and argue that not only the existence of a 1PP is constitutively necessary for the occurrence of a mind, but that anything given from a 3PP may at best be materially necessary for a mind. The exploration of this possibility would lead us too far astray, into the colossal literature on inverted qualia, zombies, disembodied minds, and even panpsychism. Luckily, it is not of particular interest for present purposes, so it will be best put aside, and I will content myself with the conclusion that, were cognition not to exist, then an account of the mind, i.e. of the target phenomenon, would unavoidably have to acknowledge the necessity of the 1PP.

What about the other, opposite assumption? Suppose that the previous assumption was false, that there indeed is a legitimate, robust, natural phenomenon corresponding to what we presently call cognition (if using the term in a rather vague manner). Let us further assume that there is no commitment concerning whether cognition has to relate to consciousness in one way or another. It may be the case, but it may as well turn out not to. This is to avoid any question-begging moves, as consciousness, just like mentality, does seem to require the existence of a 1PP, for it seems absurd to maintain that there can

be consciousness without it being witnessed from an internal perspective in any way. In the next two sections, I will investigate whether there is any reason to think that the 1PP is a necessary component of cognition to incorporate in a good MOC, in addition to the epistemic remarks concerning the "intelligibility" of what cognition and cognitive subjects are.

2. On the difference between 1PP and 3PP

What in this and the next section I will try to argue for is that, under the assumption that cognition really is something, i.e. it exists in a more robust manner than as a fictional theoretical entity, the 1PP cannot be expunged from an account of cognition. This is because, if the assumption is correct, the way the 1PP would be necessary for cognition, presumably, would not be restricted to the important but less substantial matter of how we are to achieve any understanding of it. That is to say, what I will try to do is to provide a "metaphysical" argument for the necessity of the 1PP for cognition, rather than an epistemic one as that presented in the previous chapter, according to which we would lose our conceptual grip on what the target phenomenon is. The first part of this metaphysical argument for the necessity of the 1PP, with which I have been concerned in the last section, was straightforward. If cognition turned out not to really exist, then the only phenomenon worth targeting, whose existence cannot be sensibly questioned, would be the mind. But a mind must have a 1PP associated with its activities, so an account of "cognition-turned-out-to-be-the-mind" would require involving the 1PP. The second part of the argument is going to be more complex, and before properly addressing it (I will do it in the next section), it will be important to explain as precisely as possible what the main features of the 1PP and the 3PP I will be concerned with are.

In order to show that the 1PP is metaphysically necessary, and not just epistemically indispensable, one needs to preliminarily show that it is not just an angle from which cognition is observed, but something more substantial than that. That is to say, one needs to show that the difference between something being given from a 3PP or from a 1PP is not analogous to the difference existing between observing the same object with naked eyes or through the use of some instrument which, say, relies on means other than visible light to gain access to the observed object. To make an example, consider an X-ray

machine. This machine allows us to detect the presence, shape, and other properties of the bones, muscles, and other things that can be found inside someone's body. Those things can also be observed directly, with one's own eyes, if the circumstances are appropriate, for instance if they are sufficiently big and if one were to cut someone's body open. Hence, using X-rays or one's naked eyes does not make any difference in terms of the sort of relation between the observer and the observed object. To be sure, using technology allows observing the objects inside a body without having to cut the body open, but aside from this and other practical advantages, there is not much difference between the two methods.

Now, this absence of more profound differences is precisely what we need to show *not* to obtain when cognition is observed from a 1PP rather than from a 3PP. One way to go about it is by showing that the 1PP is not reducible to the 3PP, so that they are radically different sorts of things, which imply different facts. This can perhaps be done by pointing at the properties that the two perspectives do not share. The best candidate in this sense is the fact that the 3PP is the perspective one can have on something *qua* distinguishable entity from that something, contrarily from what is the case for the 1PP. In other words, in *x*'s having a 3PP on *y*, a distinction between *x* and *y*, which may not be numerical in nature, is presupposed. This, in turn, means that having a 3PP on a cognitive system, even when it is the system itself occupying the position of the observer entertaining this perspective, can be understood as having a higher-order, thematizing perspective, where the nature of the object accessed in this way is given as something discernible, and meaningfully separable, from the act of accessing it. This is a point reminiscent of the difference between first-order and higher-order thoughts[23]: a thought and a thought about that first thought are not the same thing, if they are to be individuated in terms of their content. A first-order thought is not meaningfully separable from its content, that is, from what is being given by entertaining it, because there cannot be empty thoughts. On the other hand, a second-order thought, i.e. a thought about a first-order thought, cannot exist as a thought without the object it has as its intentional content, but it can be conceptually separated from the content of the thought it is about. For instance, having a sensation of pain cannot be meaningfully divided into two components, namely the pain and the

[23] Here, I am using the term "thought" in a very broad sense, also understandable as equivalent to "mental state": a feeling of pain is, broadly speaking, a thought, just like an emotion or a propositional thought are both thoughts in this sense.

corresponding perceptual state. On the contrary, the thought "I have a pain in my leg" can be discerned, as a thought, from the painful sensation itself. Sure, for the thought "I have a pain in my leg" to have the content it has, and consequently in order to exist (if we assume that thoughts devoid of any content do not exist), the sensation of pain must exist as well; but it is not a non-sense to distinguish the thought about one's sensation of pain and the pain itself.

As it happens, while thoughts in general are individuated by their content, so that a thought essentially is what that thought is of[24], higher-order thoughts do not appear to be reducible to the content of the thoughts they are about. At least in the case of thoughts, reducibility is not a transitive relation. The reason for the intransitivity of the reducibility-to relation in this domain is that the intermediate link, the first-order thought (which is the object of the higher-order thought and which mediates between this higher-order thought and its own content), does not cease to exist, when it is not the object of a higher-order thought. Rather, it exists as a properly entertained thought. On the other hand, that first-order thought only appears, so to say, "bracketed" or, more properly, thematized, when it is the object of the higher-order thought.

What I want to suggest is that something similar happens when one entertains a 3PP on something. To have a 3PP is to thematize what the 3PP is being had on, to pose that thing as an object to extrinsically interact with. Hence, something containing third-personal information about something else necessarily has to be a distinguishable entity from the relevant object. Just like that famous artwork by Magritte brings to our attention that a painting of a pipe is not a pipe, a mathematical model of some mechanical dynamic process is not the mechanical process itself, a mental image of a sunset is not that sunset, and a mental state M_1, which is about another mental state M_2, is not M_2. In all these cases, lots of information concerning the thematized object is involved, but there is a clear distinction between the entity carrying such third-personal information and the event or entity that information is about. Furthermore, this distinction is not just an epistemically important one, but it is also, and crucially, a metaphysically relevant one. The reason why, as a metaphysical distinction, this is a crucial one is that, were it not in place, it would not

[24] Although not in the sense that a thought is identical with the reference of its intentional content. A perception of a tree is individuated by what is being presented as perceived, but, as a thought, the relevant perceptual state is not individuated by the tree in the external world. Of course, naïve realists about perception would disagree.

be possible to have access to anything other than what whoever benefits from that accessibility has thereby access to. This seems to hold, for instance, to mental phenomena broadly construed. In other words, and straightforwardly applying this point to the specific case of cognition, had there not been the possibility to have a 3PP on cognition, and had the 3PP not been working, in general, in the suggested thematizing way, no access to others' cognition would be possible. The epistemic role played by the 3PP is only possible in virtue of its metaphysical features.

Now, the 1PP works very differently from how the 3PP works. Given the immediacy of the access to the information borne by the structures on which the subject has a 1PP, the "distance" that exists between the subject and some information-bearing structure when the former has a 3PP on the latter disappears. Hence, to have a 1PP on something is to have access to that phenomenon in a non-thematizing way, that is, while not positing that something as "other" from the entertainer of the 1PP perspective. To entertain a thought, to have a 1PP on a thought, is to be the thinker of that thought. To put it in a stronger way: to have a 1PP on a thought is to *be* that thought. There is no distinction between whoever entertains a thought, *qua* thought-entertainer, and the thought which is thereby being entertained. One's thoughts are being lived through as they occur, and their subject is not something over and above them; rather, the subject of a thought is something which, from a 1PP, is built in the thought itself. A subject of a thought is not something which entertains some relation with their thought(s) as one would entertain it with an external object: ownership of a thought is not the same sort of ownership one can have with respect to some physical entity in the world outside one's body.

Given the way the 1PP works which I have just presented, then, one can only have a 1PP on one's own thoughts, as one does not exist as a sort of substrate underlying them, so that whenever there is a thought with a 1PP associated with it, whoever is having that 1PP will also be the legitimate subject, or entertainer, of that thought. This much seems particularly evident in the case of consciousness. But my target is not, at the moment, consciousness. Indeed, it will be worth pausing a moment to highlight some similarities and differences existing between the notion of the 1PP as I have characterised it up to this point and other notions from the literature on consciousness This is because some of these notions are in the conceptual neighbourhood of the 1PP, and they may legitimately come to one's mind in trying to understand my proposed definition of the 1PP. Specifically,

one may wonder whether the definition of the 1PP means that, for something to be given via a 1PP, it is to be available to introspection, and whether to have a 1PP associated with some cognitive phenomenon amounts to that phenomenon's being access conscious.

Let us start from the former possibility, namely that to be given via a 1PP is to be available to introspection. Introspection is typically taken to be a process that has to do with the intentional creation of some higher-order cognitive state whereby the presence and content of some other cognitive state is acknowledged (for an overview of the generally accepted features of introspective processes, see Schwitzgebel (2019)). For instance, I may become aware of being attracted to someone as the result of an introspective process if I form the relevant belief as a result of focusing my attention on the feelings that the other person's current presence sparks in me. In this case, I would form a belief, "I am attracted to *x*", whose content is about some other mental state of mine (the feeling of attraction to *x*).

Based on the given definition of 1PP, it seems that, at least *prima facie*, in a scenario such as the one just sketched I could be said to be entertaining a 1PP on my feeling of attraction: the content of whatever information-bearing structure conveying my feeling of attraction is made available to me, the relevant cognitive subject. Not only that, but if such information-bearing structure has nonderived content, then this content needs to be made available to me in an immediate way, as per my understanding of nonderived content. However, to maintain that would be a mistake, given my characterisation of the 1PP and of the 3PP. First of all, if one conceives of introspective acts as involving the formation of some higher-order cognitive state, then introspective acts only afford a third-personal, rather than first-personal, access to the introspected states. This is because the formation of a higher-order state unavoidably consists in thematising the relevant lower-order state. That is to say, an introspective state may make the content of the introspected state available to the subject, but it does so in a way that does not depend exclusively on the interaction between the introspected state and the subject; rather, this availability is achieved through the mediation of the introspective state itself. In this sense, if introspection involves the formation of higher-order states, then it does not make the content of the introspected state *immediately* available to the subject, as the definition of 1PP instead requires.

On the other hand, if one does not require that (all) introspective acts rely on the formation of a higher-order state, identifying the 1PP with introspection would still not be appropriate. This is because, under such a reading of introspection, the subject would still need to become somehow aware of the content of the introspected state, but no such awareness is required by the definition of 1PP. It is in fact entirely possible that the content of some information-bearing structure is made (immediately) available to the subject without the subject being aware of it: for some content to be made available to the subject it is for that content to affect the cognitive activities of the subject, but this may happen unbeknownst to the subject itself. In other words, one may maintain that introspection entails the presence of a 1PP in my terms, but not vice versa. Not all first-person-perspectival phenomena are available to introspection.

What about access consciousness, then? It seems that what I have just said about introspection and the 1PP naturally leads towards an understanding of the 1PP as access consciousness (Block, 1995), where for some content to be access conscious is, roughly, to be poised for rational processing, without the need for the subject to be aware of this fact. Based on Block's characterisation of the notion of access consciousness, it is indeed plausible to understand the 1PP as extremely close to that of access consciousness, but with three caveats. First, the 1PP is meant to be a notion related not to consciousness, but to cognition more broadly construed. Insofar as access consciousness requires that the relevant content is made available for conscious (even though, perhaps, phenomenally unconscious) cognitive activities, it follows that not all processes that have a 1PP associated with them are access conscious. This is because the impact, had by the availability of the information associated with the relevant information-bearing, may not manifest itself in a conscious way (indeed, the cognitive system in question may not even be capable of consciousness at all).

Secondly, my definition of 1PP states that the content of the relevant information-bearing structures is made available immediately. Access consciousness does not pose such a constraint. It is thus in principle possible that, in my terminology, one is access conscious of some piece of information only in a third-personal way, but not in a first-personal way.

Third, and most importantly, the notions of access consciousness and of the 1PP as I have characterised it here may be very close to one another, but they are not equivalent.

For, it is possible that there is a 1PP without access consciousness, at least insofar as it is possible there to be phenomenal consciousness without access consciousness. Suppose that for some period you are in a gloomy mood. To be in a gloomy mood certainly has a phenomenal component to it, and it affects the cognitive phenomena you display (behaviours, thinking, feelings, perceptions...) even while you are not access conscious of being in that mood, say, because you are focused on some activity. In this sense, the phenomenal content of the neural configurations could still be made available to you as a subject without you being access conscious of it. And, if such availability is immediate, this means that there can be a 1PP associated with your mood in the absence of access consciousness of it.

In short, my conception of the 1PP undoubtedly has some affinity with the more familiar notions of availability to introspection and of access consciousness, but it also differs from them under some respects. With these clarifications under our belt, it is time to move to the argument for the necessity of the 1PP for cognition.

3. Why having a 1PP may be necessary for cognition

At last, in this section I will provide an argument to the effect that cognition cannot exist without there being a 1PP associated with it. As I have said, such an argument should not be an epistemically driven one, to the effect that we cannot conceive of what cognition without a 1PP would be like. Rather, this argument has to be one to the effect that without a 1PP there indeed could not be cognition, meaning that cognition could not exist at all, regardless of whether anyone is capable of understanding it or not.

If cognition is a really existing natural phenomenon, then it should be something which can exist independently from any (external) observer. This, obviously, does not mean that it can exist independently from any mind (after all, in a way it is of "the mind" that we are talking about at the moment). What I mean, instead, is that something could have cognitive status, i.e. should instantiate cognition, even if no 3PP is being had on it, in the sense explained above where having a 3PP on something consists in entertaining an extrinsic, observer-object relation. But one may wonder: is it possible to preserve something's cognitive status while no 1PP is being had on it? Is it possible for cognition

to exist without any 1PP being associated with it, or is the existence of such a 1PP necessary for any cognitive activity to happen?

In order to show that *x* is necessary for *y*, one needs to show that *y* cannot possibly occur without *x*. If one has a clear conceptual grip on *y*, or, even better, a definition of it, this can potentially be shown conclusively, as what is needed is a proof that the properties defining *y* cannot occur without the occurrence of (the properties of) *x*. Indeed, the modal component of the correlation between *x* and *y* depends on this. But how can this be done if a definition of *y* is not available, as is the case for the claim that 1PP is necessary for cognition, given that the whole point of the present inquiry is going towards a currently missing definition of cognition? If one only has an intuitive, implicit, largely pretheoretical characterisation of *y*, i.e. of cognition in our case, it is just not possible to immediately add the modal component. Consequently, if one has to find the necessary conditions for something which is, in and by itself, vaguely understood, the best one can hope for, at first, is to find something that systematically correlates with *y*. Or better, something whose absence systematically correlates with the absence of the target phenomenon. Then, the necessity claim can be added by providing some satisfactory enough explanation for this correlation. The problem for the case at hand is that it is not just the target phenomenon that is vague and far from being successfully captured by a definition, but also the candidate necessary condition is difficult to pin down and spell out exhaustively. Not only that, but, as if that were not enough, the *x* of the relation, i.e. the 1PP in the present case, is something which, by its nature, cannot be accessed by anyone who does not entertain it, thus by anyone who is not the subject of the target phenomenon or, one may say, by anyone who is not a particular instance of the type of target phenomenon itself.

However, these considerations exposing the difficulties of the task I will undertake here only have the consequence of preventing the argument from being conclusive. This does not mean that it cannot be persuasive enough or reasonably strong. Therefore, let us now turn to the argument itself, and let us try to make a satisfactory case for the necessity of the 1PP for cognition.

To lack a 1PP is to lack a subject for the relevant cognitive phenomenon, in the thicker sense of subject which relies on a somewhat *sui generis* constitution relation. This point will be best illustrated by means of a (dis)analogy. Consider the example of a striped

flag. The coloured stripes in a flag belong to the flag, in the sense that they constitute it for what it is, but the flag is not the subject of those stripes. The stripes contribute in an indispensable way to the flag's occurrence as the flag it is, but, putting semantic perplexities aside, the flag is not the owner, nor the subject, of the stripes. Constitution is instead associated with subjectivity and ownership in the case of cognition. A cognitive system, if it is a proper one, is constituted by its cognitive states and processes, and it is their legitimate subject as a consequence. Had the relevant cognitive states and processes not occurred, our hypothetical cognitive system would not have been such in the same way it is, and perhaps it would not have been a cognitive system at all.

The question then becomes: can a cognitive system be a cognitive system without also being a cognitive subject in the sense warranted by having a 1PP associated with cognition? Intuitively, one would (or at least, I would) say no. But for what reason? To see this, suppose that I claim to be the owner of the cognitive states *C* of a hypothetical system *S* without there being a 1PP associated with them. If it were possible to have cognition without simultaneously having a related 1PP, the fact that I do not have first-personal states associated with the operations with (we are supposing) cognitive status in question would not constitute an objection to my claimed ownership with respect to those cognitive states and processes. My claim would not have to be read in terms of being a subject of that cognitive activity, since by assumption such activity has no subject, i.e. it has no first-person-perspectival states associated with it. Rather, my claim would be analogous to the claim that something is part of something else. The process going on in the alleged cognitive system *S* would be part of the overall cognitive economy of the extended system "me + *S*". And how is one to establish whether this parthood relation obtains? One way to address this issue that seems appropriate for the present framework is by arguing that, if some appropriate causal coupling is in place, then *S* is part of the system "me + *S*"[25]. Presumably, such coupling would have to be understood in functional terms, for instance by saying that the information processed by *S* is made available to the rest of the system, in one way or another.

This last point seems to be remarkably close to what Rowlands proposes in saying that the information processed by a cognitive process has to be made available at the

[25] I am not interested here in discussing mereological questions that would arise from this. Parthood here is to be understood in a simple way, in a sense similar to what being a part of a functionally characterised system means.

personal level, either directly or by being made available to subsequent processing first. The problem is that there is no way to demarcate the boundaries of a cognitive system if this were the only criterion. Indeed, in his proposal for a MOC, Rowlands makes a crucial use of the notion of subject: the personal level is to be determined with respect to the cognitive subject, and the cognitive process has to belong to this subject. Or, more precisely, the representations (with nonderived content) by means of which the information involved in the process is (meant to be) made available to the subject have to belong to the subject itself. But in the present scenario there is no subject involved, as I have argued that the existence of a subject depends on the existence of a 1PP, and there is no 1PP by assumption.

So, why could ownership of the cognitive states *C* of *S* not be arbitrarily claimed? The information involved in cognitive processing is plausibly Shannon information (i.e. the understanding of information as reduction of uncertainty). But Shannon information is importantly determined by the receiver's predisposition to receive and extrapolate such information by causally interacting with the relevant information-bearing structures. Therefore, any entity capable of exchanging information in this sense with some other entity performing a putatively cognitive task would have a legitimate claim to the ownership of the related cognitive processes. This is because the information cognitively processed would be accessed by the first entity, and that would meet the functional requirement mentioned above.

The resulting scenario would be analogous to the one involved in the cognitive bloat objection which critics of the Extended Cognition view used against the possibility, or at least of the actual existence nowadays, of extended cognitive systems. The gist of the cognitive bloat objection, in the context of the Extended Cognition view, is that as soon as one admits that a paradigmatic cognitive system such as a human being (or perhaps just their brain) can be part of an extended cognitive system composed of external entities, which are as legitimately constitutive of the overall system as the human is, a dangerous slippery slope cannot be avoided. If a mathematician's calculator is conceded cognitive status as an important component of the "mathematician + calculator" cognitive system, then where should the line be drawn? Why could the mathematics books which the mathematician studied during her degree not be included? After all, they importantly played a role in the calculations being currently performed. Or, to push the absurdity

which the cognitive bloat aims at highlighting even further, why would intuitively mere background conditions, such as the existence of electromagnetism in our world, be ruled out as an important part of the cognitive system at issue? Without electromagnetism, the cognitive performance in which the mathematician with her calculator is engaged could not take place. In short, the cognitive bloat objection is meant to show that if too loose requirements are adopted for being a constitutive part of a cognitive system, then it is hard to prevent too many things from being included, in an undesirable way, in the cognitive system[26].

In a somewhat similar spirit, if the only requirement for a cognitive state to belong to a cognitive system is that it exchanges information with that system, then any system could encompass any information-bearing cognitive state, as there virtually is an exchange of information (perhaps a very indirect, even unsuccessful one) between any two things in the world. Ruling out the existence of a cognitive subject in the sense afforded by the presence of a 1PP would prevent us from introducing tighter requirements such as Rowlands's criterion, according to which for a process (or state) to qualify as cognitive, the information which it involves has to be made available, at some point, to the cognitive subject by means of representations owned by it.

Furthermore, not only the scenario obtained would be troubled by the analogous of the cognitive bloat objection, but the assumption under which we started working, namely that cognition really exists, would be jeopardized. The reason for this is that if some activity qualifies as cognitive, and if cognition is something which really exists, i.e. which is not a mere instrumental fiction consisting in a relational, observer- and theory-dependent quantity, then whether and how, at any given time, cognition is displayed by something should not change, all things being equal. But this is not the case in our hypothetical scenario. The cognitive system *S*, coupled with me (the claimer of ownership over the processes *C* of *S*), would contribute to the overall cognitive economy with a certain amount of information, but such amount would be different in different contexts. Furthermore, all such couplings ("me + *S*", "*x* + *S*", …) would be somewhat arbitrary, and unavoidably so, in the absence of a criterion sufficiently robust to not leave the boundaries of cognitive systems underdetermined. However, this criterion seems to be

[26] Incidentally, this is one of the reasons why elaborating as precise a MOC as possible is important.

only obtainable if the notion of cognitive subject is in place (and, in our hypothetical scenario, it is not in place, given the absence of a 1PP).

In the previous chapter, I discussed the possibility that the choice of cognitive systems is under-constrained by a plausible necessary condition, and I have concluded that cognition understood in that way, i.e. as mere Shannon information, is not theoretically useful, especially in the sense that it would not be enlightening enough for us. Here, I went further and pointed at the fact that that weak conception of cognition, which does not require the existence of a 1PP, would not only be scarcely useful from a theoretical point of view, but it would also correspond to nothing in the real world. Just like there is no natural property corresponding to information in Shannon's sense, there is nothing in the world which, independently from any theory, corresponds to cognition. Therefore, we appear to have reached a contradiction. If we work under the assumption that cognition is not a theoretical fiction, then also assuming that there is no 1PP on it leads us to the conclusion that, after all, cognition is not a really existing natural phenomenon. Then, there are two possibilities. One can drop the initial assumption that cognition exists in a robust sense; but this would bring us back to the first part of the argument, so that the phenomenon we were trying to capture was the mind all along, and since minds necessarily have a 1PP associated with them, an account of "the-cognitive-turned-out-to-be-the-mental" would have to necessarily involve the 1PP. Or one can simply drop the second assumption, and accept that not only cognition exists, but there has to be a 1PP associated with it.

In summary: either the 1PP is not necessary for cognition, in which case cognition is not a genuine phenomenon to account for, or the 1PP is necessary for cognition. Therefore, since cognition *is* a real natural phenomenon, the 1PP should be taken to be necessary for cognition.

Concluding remarks

In a way, there is some affinity between the view outlined in this chapter and Rowlands's (2010) view, which I have touched upon at the end of the first chapter. In Rowlands's view, cognition is a disclosing activity that requires a subject. Similarly, the 1PP is the perspective that a cognitive subject has on its own cognitive activity, and in

this sense, claiming that the existence of a 1PP is necessary for cognition is equivalent to the claim that cognitive processes must have a subject, something that Rowlands stresses in a number of places[27]. Furthermore, according to Rowlands's second condition, cognitive processes have the (proper) function of making available to the subject the information they process, which is something embedded in the definition of the 1PP. The crucial difference between my proposal and Rowlands's consists in the fact that I maintain that there must be no mediation. In my view, representations with derived content cannot constitutively take part in a cognitive process, because their content is not immediately available to the subject, while in Rowlands's view they can, as long as the relevant information is ultimately made available to the subject. I suppose that the disagreement here ultimately boils down to a difference in intuitions. Think of scenarios coming from the phenomenological tradition, where a blind person ceases being aware of the cane and just is aware of the information that is being processed in their using the cane[28], or where a person engrossed in reading a novel stops being aware of the written words as words, and just is aware of their meanings. My intuitions tell me that the medium doesn't cease to be a medium just because one stops being aware of it. Proficient, effortless, and even apparently "transparent" tool use is not a symptom of the disappearance of a distance between the tool and its user. Rather, I find it more plausible to say that the user stops being aware of such distance, which nevertheless is not bridged. This is where my intuitions appear to differ from Rowlands's. For a cognitive process to belong to a subject, it is not enough, in my view, that the subject is not aware of the fact that there is a distance between the contents of its mind and the content of certain portions of the process. Such distance must cease to exist in a way that does not depend solely on the subject being oblivious to it.

Of course, there is plenty of room for objections if one shares Rowlands's view. Luckily, this disagreement happens to be of little import to the present work. The 1PP proposal seems to be in line with Rowlands's view at a general level, that is, insofar as it is not applied to particular cases, be they thought experiments or real cases. But in the next two chapters, I will offer reasons to be unhappy with the proposal in the first place, independently from its application to particular cases.

[27] For instance, 'there are no subjectless cognitive processes' (2010, p.140).
[28] A scenario originally appearing in Merleau-Ponty (1962).

Chapter 4

Introduction

In this chapter, I will develop one of the two lines of criticism that will lead, in the next chapter, to the rejection of the requirement that all cognitive processes must have a 1PP associated with them. I will argue that there is reason to think that the 1PP condition implicitly sneaks consciousness into the picture, so that if one were to accept it, then one would plausibly also have to accept that there can be no unconscious cognitive phenomena.

I will proceed as follows. In the first section, I will show that there is a significant overlapping between the notion of the 1PP I have been discussing and the notion of the 1PP as it figures in some of the contemporary phenomenological literature. Since the latter is taken to be an ineliminable component of conscious experiences, one may wonder whether this means that the phenomenon with respect to which the 1PP may be taken to be a necessary component is not cognition, but rather consciousness. If that is the case, then taking the 1PP condition on board would suggest that one is also committed to the further claim that there is no such thing as unconscious cognitive processes. This is because the sort of cognitive processes for which the 1PP is necessary would be *conscious* cognitive processes. And, given that the 1PP is meant to be associated with all the occurrences of cognitive processes, the existence of unconscious cognitive processes would be denied.

This implication is a questionable one, and in the second section I will present a phenomenon which would constitute a straightforward counterexample to it, namely somnambulism. Since sleepwalkers are occasionally capable of engaging in activities which should plausibly be regarded as cognitively driven while apparently being unconscious, it must be possible for cognition to occur in the absence of consciousness. Of course, somnambulism is by no means the only phenomenon which appears to involve unconscious cognitive activities. Nevertheless, it is a striking exemplar, and, importantly, one which does not seem to afford an easy way out for those who may want to address it by disqualifying the relevant activities as non-cognitive.

Although one cannot dismiss the somnambulism counterexample by claiming that sleepwalkers never engage in cognitive activities, perhaps one can defend the 1PP condition by biting the bullet. That is, the 1PP condition could be salvaged if one were to find a theory of consciousness able to block this objection, by showing that somnambulists manifest at least some form of minimal consciousness. I will then discuss in the third section a theory of consciousness which seems prima facie capable of delivering such a result, namely the Integrated Information Theory (IIT) of consciousness. This theory has been chosen among the vast number of theories available not only because it is one of the most popular at the moment, but also because there is a tendency among its upholders to gather evidence in favour of the theory itself from empirical studies on NREM sleep (when somnambulism occurs) and somnambulism. Thus, it is arguably the best theory of consciousness (or, as a minimum, among the best ones) in terms of its ability to shed light on the possibility that somnambulism involves consciousness, at least to some degree.

Finally, in the fourth section I will show that, despite its potential, the IIT ultimately proves unsuitable as a theory of consciousness for the purpose of preserving the 1PP condition. This is for two main reasons. First, its axiomatic guise is widely criticised as unsatisfactory. Second, the computational intractability of integrated information prevents the theory from being appropriately empirically tested. As a consequence, despite being in principle a suitable candidate, the IIT is not at present sufficiently robust to defend the 1PP condition from the somnambulism objection. This, however, will not be a lethal attack against the 1PP condition, for it relies on the assumption that the condition does imply that there cannot be any unconscious cognition. For this reason, the next chapter will develop an alternative line of criticism against the 1PP condition.

1. The 1PP and consciousness

Let us start by recalling the notions of 1PP and 3PP. Cognitive phenomena seem to be observable from two different points of view. The first (3PP) is the perspective which an external observer can assume, and the second (1PP) is the perspective which the subject of the phenomena themselves can assume "from the inside". Observing, describing, and understanding cognition from the 3PP typically means that one has access

to what cognition looks like in its material manifestations. One can study cognitive systems by observing their material realisers, describing their components and the causal processes they instantiate, and understanding the dependency relations that underlie the different activities that they engage in. Hence, having third-personal access to cognition means having access to what cognitive systems are made of, to the outcomes and patterns of the behaviours they display, and, at best, to the underlying causal mechanisms that make their occurrence possible[29].

On the other hand, the 1PP on cognition is the perspective that the cognitive subject itself has on its own (first-order) cognitive goings-on. One could say that the 1PP of a cognitive subject consists in metaphorically witnessing the subject's own cognitive states and processes, but this could be misleading, as not every cognitive state or process corresponds to another cognitive state which is about the first (i.e. a higher-order cognitive state or process). This would require a form of self-awareness that does not seem to be displayed by every putative cognitive system: when an animal is hungry, it is not at all obvious that it thinks "what I am feeling is hunger", not even in non-propositional form, because that would require the quite sophisticated sense of the self. Similarly, when someone, human or not, sees something in front of themselves, it is not necessarily the case that a conscious thought along the lines of "there is something there" is entertained, be it linguistically formulated or not. A perhaps more acceptable way of characterising the 1PP is saying that the 1PP on a cognitive process consists in *being* someone, and more precisely, in being that someone to whom that process belongs. Or, better still, assuming that a cogniser is nothing over and above the totality of its cognitive activity, one has a 1PP on a cognitive process if one *is that cognitive process*, with all the consequences that this has: experiencing the feeling of being so cognising, if there is any such feeling; having no (pre-linguistic) doubts that that process belongs to oneself; knowing without reflection what that process appears to be from the inside.

In the previous chapter I have explored the view that the 1PP may be necessary for cognition. It is important, however, to be clear about what sense of "being necessary for" is exactly at play here. In the context of the inquiry on the nature of cognition, there are,

[29] To be precise, in the previous chapter I have argued that one can have a 3PP on a cognitive process also while being in a higher-order cognitive state. That is, one can be in a first-person-perspectivally given cognitive state *x* which carries information about another cognitive state *y*, and thereby have a 3PP on *y*. This sort of instances of 3PP are not important for the present chapter's purposes.

broadly speaking, two senses in which the claim that some state of affairs or entity is necessary for a given phenomenon can be understood. The first has to do with what constitutes the target phenomenon itself. Under this reading of necessity, something is necessary for the target phenomenon in that it is part of what that target phenomenon is. For instance, the white cross of the Scottish flag entertains this relation with the Scottish flag, for the Scottish flag could not be the flag it is, were it not to display that cross. Or, to use another example that does not have to do with parthood relations, for some substance to be a metal, that substance needs to be a good electrical and thermal conductor (among other things). Or again, for a process to qualify as a photosynthetic process, the transformation of light into chemical energy needs to occur. In these examples, the mentioned conditions are necessary for the object or process in the constitutive sense: they are part of what it is for the thing in question to be the sort of thing it is.

The second sense of necessity has instead to do with the enabling conditions for the occurrence of the target phenomenon. In this sense, a necessary condition is not an essential aspect or component of the phenomenon itself, but it is something that plays a crucial role in bringing the phenomenon about. For instance, a necessary condition in this second sense for something to be a footprint is being causally brought about by contact with a foot. Or, for a child to be born, it is necessary that a fecundation occurred.

Undoubtedly, the distinction between these two senses of necessity is not a sharp one, and, in most cases, it will be possible to interpret a given necessity claim as an instance of either of them. Nonetheless, it is the first of the two senses of necessity that is relevant for the claim that the 1PP is necessary for cognition. The existence of a 1PP does not diachronically bring cognition about. Rather, having a 1PP associated with it is partly constitutive of what it is to be a cognitive process. With this clarification in mind, we can return to a claim that I have mentioned more than once in the previous chapters, namely that a satisfactory account of cognition should not neglect the 1PP. If the 1PP is necessary for cognition in the sense that it is constitutive of it, then the better an understanding of the 1PP we have, the firmer our grip over the notion of cognition will be, as the 1PP is part of what cognition is. This is because, insofar as the presence of a 1PP is a constitutively necessary condition for cognition, the 1PP itself is part of the explanandum. Differently from what happens in the case of "enabling" necessity, our understanding of

the very nature of the target phenomenon cannot be separated from the nature of the proposed necessary condition.

Let us then rehearse what has been said about the nature of the 1PP so far, for this will shed light on (part of) the nature of cognition. The 1PP is the non-thematizing perspective that a cognitive subject has on its own cognitive activities, and it is such that the information accessed through this perspective is given to the subject immediately and in an exclusive way (i.e. only the subject of the relevant cognitive process can have first-person-perspectival access to the information processing in question). Based on this characterisation, one may wonder about the relation that may exist between the 1PP claimed to be constitutively necessary for cognition and consciousness. Consider the following passage from Zahavi and Parnas (1998, p.689):

> […] in much phenomenological literature, the discussion of self-awareness […] is a discussion of how consciousness is aware of itself. In other words, the question of self-awareness is basically taken to be a question of how consciousness experiences itself, how it is given to itself, how it manifests itself.
>
> […] we will speak of the first-personal givenness of phenomenal consciousness in terms of self-awareness. Whereas the object of my perceptual experience is intersubjectively accessible in the sense that it can in principle be given to others in the same way that it is given for me [in a 3PP fashion], my perceptual experience itself is given directly only to me. It is this first-personal givenness of the experience which makes it subjective.

There is no doubt some striking similarity and overlapping between the way I have characterised the 1PP in the context of the study of cognition and the way contemporary scholars in the phenomenological tradition characterise it in the context of the study of consciousness. Not only that, but in the first chapter I have explicitly drawn inspiration from the "self-referential" character of consciousness for the claim that nonderived content needs to be somehow tightly and intimately related to the subject of the relevant representations. The aforementioned worry then becomes a substantial one: what if the aprioristic line of inquiry adopted thus far resulted in a concealed shift of the focus of the inquiry itself, from cognition to consciousness? If the 1PP is indeed constitutively necessary for cognition, and if it also is a core feature of (phenomenal) consciousness, is it not possible that the constitutive necessity claim explored in the previous chapter

ultimately leads to the conclusion that cognition and consciousness are one and the same thing? Of course, similarity and overlapping of certain features are not enough to conclusively establish an identity between what possesses such features, at least in most cases. But, to use legal jargon, circumstantial evidence calls at least for an alibi: given that the claim that the 1PP is necessary for cognition ultimately stemmed from considerations based on intuitions and introspection, and given that we are conscious creatures, it would not be too speculative to suppose that the pre-theoretical understanding of the target phenomenon steered the inquiry in the direction of consciousness proper, rather than of general cognition.

The first step to take is to examine what the implications of this worry are. Suppose that, based on the given characterisation of the 1PP and on the claim that the 1PP is (constitutively) necessary for cognition, it really should be concluded that cognition and consciousness are not distinct phenomena. Then, it would immediately follow that there is not such a thing as unconscious cognition. This conclusion would, of course, fly in the face of current scientific consensus. But how damning for the view in question would this fact be? It is important to keep in mind the context in which such a conclusion would be drawn. The disciplines falling under the umbrella of the cognitive sciences have achieved significant empirical and theoretical robustness. Nonetheless, even though this is unlikely to have too negative an impact on scientific practice, there is at present still a non-negligible gap in our knowledge where an accurate characterisation of the target phenomenon, cognition, should be located. In other words, the definition of what cognition is, is still up for grabs. This means that, even if this would plausibly not affect the way most of the cognitive sciences are practised, there still is room for potentially revisionary proposals concerning the nature of cognition, even for consciousness-centric ones. The empirical achievements of the cognitive sciences would not be tarnished were it to emerge a consensus around the idea that much of what the cognitive sciences are concerned with is not really cognition, but something in its conceptual neighbourhoods, directly or indirectly related to cognition.

The upshot is that the fact that there presently is ample consensus to the effect that unconscious cognition can and does exist is not a conclusive reason against the picture outlined up to this point. This is because, although the disciplines in question are empirical, the tension would not arise at the empirical level, but at the theoretical one: it

would have to do not with the observable phenomena, but with the way such phenomena are to be construed. Otto makes use of his notebook in exactly the same way regardless of whether, theoretically speaking, the "real" cognitive processing happens entirely within his brain or body, or whether it also constitutively encompasses the notebook. Or consider the phenomenon of masking, whereby some perceptual input is not consciously registered by the subject, but is nonetheless capable of informing some of the subject's subsequent cognitive activities. It is an empirically ascertained fact that masking occurs, but it is a matter of theoretical interpretation whether the information contained in the masked input ought to be considered as cognitively, but unconsciously, processed, or whether it is not to be considered processed in a cognitive sense at all. As a consequence, pointing at the fact that the cognitive sciences are currently taken to study both conscious and unconscious phenomena, and in virtue of this arguing that cognition and consciousness cannot be equated, won't do. It may well be the case that the cognitive sciences study empirical phenomena which are in one way or another related to cognition/consciousness, but which are not cognitive themselves. Therefore, as odd as the identity in question may be, more substantial reasons for rejecting it must be provided.

Where are these reasons to be found, then? On the one hand, it seems that aprioristic meditations of the sort I have been concerned with in the previous chapters led us to the view that the 1PP is necessary for cognition, and that in all likelihood cognition and consciousness are one and the same thing. On the other, empirical scientific practice takes as an established fact that unconscious cognition exists, so that either the 1PP is not necessary for cognition (because it would lead to the further consequence that all cognition is conscious), or some reason to believe that the claim concerning its necessity does not lead to the further identification of cognition and consciousness is required. Since the whole issue stems from the fact that this identity appears to naturally follow from the way I have characterised the 1PP and the way certain distinctive features of consciousness are characterised, the latter alternative is not viable. Therefore, it seems that a *prima facie* impasse has been reached: pure speculation leads us to one conclusion, but empirically informed speculation leads us to another.

To settle the matter, the best course of action is to explore the feasibility of the scientific theories which could be used to empirically substantiate the conclusion reached via speculation. That is to say, in order to decide not just whether there are positive

reasons to accept the aforementioned constitutive necessity claim, but also (and, perhaps, more importantly) whether there are reasons *not* to reject it, one needs to examine the state of the empirical accounts of consciousness, and in particular of those accounts that seem to best accommodate the theoretical claim in question. If the extant ones are in reasonably good shape, and if they accommodate the claim concerning the necessity of the 1PP, then one may have further reasons to accept the condition itself, even if this means accepting the identity of cognition and consciousness. This is because doing so would not put a supporter of that condition in the uncomfortable position of endorsing a view insufficiently backed by empirical accounts of the phenomenon in question (cognition/consciousness), for in that case one may be inclined to simply dismiss the view in question as the product of some misguided armchair reflections.

In short, if one can find at least one acceptable empirical theory of consciousness able to accommodate the condition on the 1PP, then the fact that this is a highly unorthodox requirement would not be sufficient grounds for dismissing it. Nonetheless, the unorthodox aura of this condition is not incomprehensible, as there exist numerous empirically observable phenomena which appear to be best accountable if one admits that not all cognition must have a 1PP associated with it. Here, I will focus my attention just on one, namely somnambulism ("sleepwalking").

2. Somnambulism

Sleepwalking is, since Broughton's (1968) seminal paper, canonically considered a disorder of arousal, i.e. a phenomenon occurring during partial arousals from the slow-wave, NREM sleep (especially during the first sleep cycle, as reported by Guilleminault et al. (2001)). Somnambulism can involve remarkable activities, which, if performed while awake, would be considered indisputably cognitively driven, e.g. cooking and eating (Schenck & Mahowald (1995)), playing instruments, driving a car (Zadra et al (2013), and talking, although, as noted by Avidan & Kaplish (2010), somniloquy is not listed under the category of traditional parasomnias and occurs most commonly during N1, N2, and REM sleep. These behaviours are not stereotypical, automatic ones, as they involve some degree of awareness of one's surroundings, despite responsiveness to external stimuli being of course importantly diminished during sleepwalking episodes in

comparison with wakefulness. Moreover, humans are typically capable of cognition under normal conditions, so there seems to be no reason to deny cognitive status to processes which are ordinarily taken to be manifestations of cognitive activity, and which are carried out by paradigmatically cognitive beings.

The problem with the view that all cognition necessarily has a 1PP associated with it, so that there ultimately can be no unconscious cognition, is that, commonsensically, sleepwalking occurs in the absence of consciousness. First of all, sleepwalkers are trivially sleeping, and sleepers are ordinarily taken not to be conscious, at least not under most readings of what "being conscious" means. Second, somnambulism is a NREM parasomnia (ICSD-III), i.e. a disturbance of sleep which occurs during the NREM phase, when the cerebral activity is significantly reduced even in comparison with REM sleep, not only with wakefulness. This latter point corroborates the hypothesis that consciousness is absent during somnambulism. Moreover, also the phenomenology of NREM sleep partially supports this hypothesis, as patients typically do not remember either their dreams or their behaviours occurring during this sleep phase (Howell (2012, p.754).

Thus, somnambulism poses a challenge for the 1PP condition, for, at least on the face of it, it appears to be at odds with the condition's implication concerning the relation between consciousness and cognition. If I am generally capable of cognition, and in particular I am thought to engage in a cognitive performance while (say) playing a musical instrument during wakefulness, why should I not be thought to engage in a similarly cognitive performance if I play during an episode of somnambulism?

This question should not be light-heartedly dismissed saying that sleepwalkers look as if they are performing cognitively driven actions, but they really are not, on account of the fact that, while sleepwalkers may be functionally equivalent to a conscious subject performing the same actions, their actions should not be considered cognitively driven because they are not conscious. Picking this line of defence for the consciousness condition would be too costly, in that, first, it would straightforwardly beg the question, as a potential counterexample would be dismissed as illegitimate simply because it would be a counterexample; and secondly, it would then become unclear whether the condition really helps understanding cognition or simply leaves the difficult cases out. Somnambulism is thus an excellent benchmark for the 1PP condition, and the way this

phenomenon is accounted for can tell us much about what the implications of accepting or giving up the condition would be.

There is certainly some sense in which sleepwalkers are not conscious. They are deeply sleeping, after all, and their unresponsiveness to attempts at waking them up, just like the confused state they find themselves in if successfully awaken, are evidence of it. However, polysomnography shows that, although importantly reduced, cerebral activity is as a matter of fact present during NREM sleep and somnambulism. For instance, Massimini et al. (2005) suggest that NREM sleep amnesia is due to a breakdown of cortical effective connectivity, but this breakdown of connectivity still happens in the context of cerebral activity. Such activity may be reduced with respect to other neurological states, but it is nonetheless non-negligible. Not only that, but some cases of sleepwalking are cases of dream enactment (e.g., the case of a father taking his infant daughter upstairs because he was dreaming that there was a fire, discussed by Pillmann (2009), and by Oudiette et al. (2009)). This constitutes evidence of mentation occurring during somnambulism, although, in contrast with a common tendency among the earliest accounts of somnambulism, it is currently accepted that thinking of somnambulism as (just or prevalently) dream-enacting behaviour is incorrect.

More generally, even though there are *prima facie* good reasons to maintain that somnambulists are not conscious, there also is some sense in which somnambulists are conscious, at least partially. This is because it would be impossible to do things such as cooking or driving without being somehow aware of the world around the agent. Furthermore, as a matter of fact, it is not entirely true that there is no conscious mentation at all during NREM sleep. While reports of dreams occurring during NREM sleep are much less frequent than those of oneiric activity occurring during REM sleep, there are some nonetheless (Cavallero et al. 1992). Besides, according to some reports, during NREM sleep dreams are closer in character to wakeful mentation than those happening during REM sleep (Kahn, Pace-Schott and Hobson (1997)), as they are not only more directly concerned with daily concerns and plans, but they also have a more "logical" structure, resembling more closely regular propositional, offline thinking than dreams occurring during REM sleep do (at least, this is true for dreams occurring during early stages of NREM sleep, according to Tononi and Koch (2008)). Consequently, NREM and wakeful mentation seem to differ more in degree than in kind (Flanagan (1995)).

If it is indeed legitimate to talk of consciousness to refer to the alertness (understood as minimal awareness of one's surroundings) required for performing the complex actions that somnambulists sometimes perform, it seems that we have roughly to do with what Block's (1995) access consciousness. Access consciousness is that form of consciousness which allows rational inferences, while it does not have any phenomenal component (although upholders of cognitive phenomenality such as Galen Strawson would disagree). This kind of consciousness can be thought of as a form of transitive consciousness, i.e. consciousness of something, as opposed to intransitive consciousness, which instead is consciousness simpliciter, the pure "being conscious" without any specific object one is conscious of. To give some illustrative examples, consciously entertaining the thought that 2+2 = 4 would be a case of access consciousness not accompanied by any phenomenal component. By contrast, being aware of the light coming from the screen of my laptop would be a case of access consciousness accompanied by phenomenal consciousness. Finally, being aware of one's own feeling of unmotivated sadness would be phenomenal consciousness without access consciousness (probably, as there is nothing that feeling is about; these are just indicative, rough-and-ready examples).

Any consciousness manifested by somnambulists plausibly has to be transitive consciousness, and more specifically access consciousness. This is because, on the one hand, somnambulists are likely not conscious simpliciter, so that they cannot be intransitively conscious. On the other hand, there presumably is nothing like undergoing whatever cognitive states they undergo, so that their consciousness is unlikely to be phenomenal consciousness[30]. The question, then, is: do somnambulists have some degree of access consciousness or not? This is the question that an empirically informed theory of consciousness should be able to address, and answer positively, if it is to support the 1PP condition.

[30] Of course, it is possible that there is indeed some phenomenal consciousness. Perhaps, upon waking up, somnambulists simply do not have any memories of that. I will for present purposes put this possibility aside, as not much hangs on whether, in addition to access consciousness, somnambulists also have phenomenal consciousness.

3. Integrated Information Theory

At present, there is a large number of attempts at scientifically accounting for consciousness. Following part of the broad taxonomy appearing in Doerig, Schurger and Herzog (2021, p.48), there are causal structures theories (e.g. Integrated Information Theory, Recurrent Processing Theory), computational theories (e.g. Global Workspace Theory, Higher-Order Thought Theory, Adaptive Resonance Theory), biological processes theories (e.g. Thalamocortical Loop Theory, NMDA Theory), cognitive processes theories (e.g. Attention Schema Theory, Sensorimotor Theory, Self Comes to Mind Theory). Reviewing all of them in order to pick the most robust one after careful confrontation would take us too far away from the object of the present research. I will therefore simply focus on one specific theory, namely the Integrated Information Theory (IIT) most notably associated with the name of Giulio Tononi (2004a, 2004b, 2005, 2008. 2012), and I will explain why it is suitable for the present purposes. This, of course, does not mean that no other theory of consciousness would do. The choice is dictated by pragmatic reasons, and, as a consequence, it will prevent my arguments from being conclusive. Nevertheless, given that much of the evidence in favour of the IIT comes from studies on NREM sleep, the choice of this theory is not undesirably arbitrary. The IIT is especially well-equipped for addressing the somnambulism objection, in that somnambulism occurs during NREM sleep. Therefore, as a minimum, should the IIT prove unsuitable for the purpose of defending the 1PP condition, the burden of finding a comparably robust and more suitable theory would be shifted to sympathisers of the 1PP condition on cognition.

3.1 Integrated information

According to the IIT, consciousness is to be understood as integrated information. That is, in a nutshell, as the amount of information that a system can generate as a whole, in addition to the information generated by its parts independently from each other[31]. Recall the notion of Shannon information standardly used in information theory, outlined in the second chapter: information is a measure of the reduction of the uncertainty

[31] I will focus here only on the quantitative part of his theory, not also the qualitative one aimed at univocally coding every particular informational state.

concerning certain states of affairs[32]. One can see that the concept of information is closely related to that of probability, as in general the more alternative possibilities are ruled out by a system's entering in a certain state, the more information is generated. However, depending on the presence and nature of causal interactions not only within, but also among a system's parts, the amount of information generated by a system as a whole can vary. In particular, if the parts of a system are not informationally isolated from each other, so that they mutually influence each other's generation of information, the system as a whole can produce more information than the sum of the information produced by its parts independently from each other. If this is the case, the system is said to produce integrated information. To see how this works, consider the two following scenarios based on Tononi's (2008) own examples.

In the first scenario, there are two switches (1 and 2), each connected to another switch (respectively, A and B). In principle, there are 16 different possible combinations of switches being on or off, i.e. the repertoire of potential states the system can be in consists of 2^4 different states, and all such states are a priori equally likely to occur. As a consequence, every state has 1/16 chance to occur, and ascertaining any overall state of the system yields $\log_2 16 = 4$ bits of information. However, the system composed of the two parts (the A-1, B-2 pairs of switches) is made in such a way that if a switch marked by a number is on, the corresponding switch marked by a letter will be switched on at a later time as well, and if the number-switch is off, the corresponding letter-switch will switch off as well. For example, if switch 1 is on, also switch A will later be on; and if switch 1 is off, also switch A will later be off. Therefore, the state the overall system is in at a certain time, t_2, narrows down the range of possible states the system was in at a previous time, t_1. For instance, if at t_2 the system is in state "(1-on, A-on); (2-off, B-off)", one can infer that at t_1 the system could have been in only 4 of the 16 potential states aprioristically considered, namely those combinations in which 1 is on and those in which 2 is off (i.e. (1-on, A-on) and (2-off, B-off); (1-on, A-off) and (2-off, B-off); (1-on, A-on) and (2-off, B-on); (1-on, A-off) and (2-off, B-on)). These four states constitute the actual repertoire of the system, and the overall state of the system under consideration carries 2 bits of information: $\log_2 4 = 2$. Now, the total amount of information is the same as the

[32] Strictly speaking, Shannon information and integrated information are two different kinds of information (as remarked by Koch (2019)). However, for present purposes, it will not be necessary to further explore the ways in which they differ.

sum of the information generated by the two parts of the system, because each part could have been in two possible states, so that each pair generates $\log_2 2 = 1$ bit of information. In this case, no information is produced by the system in addition to the information produced by its parts, so there is no integrated information.

Consider now a second scenario, in which the same switches are linked in such a way that: A can operate on 1, B and 2; 1 can operate on A, B and 2; B can only operate on A; and 2 can only operate on B. Furthermore, any switch will be on at t_2 only if, at t_1, at least two other switches operated on it. While, as in the previous scenario, each combination of switches on or off has aprioristically 1/16 probabilities to have occurred at $time_1$, if at $time_2$ A is the sole switch on, while all the other switches are off, it follows that just one among the 16 in principle possible combinations could have occurred: 1 and B were on at $time_1$, and the other switches were off. This means that the overall state of the system carries $\log_2 16 = 4$ bits of information. What about the individual parts of the system, namely the pairs A-1 and B-2? Without entering into the technical details, by applying Bayes' theorem and calculating the Kullback-Leibler divergence, one can find that the information generated by the first pair, A-1, is 1.1 bits, while the information generated by the second pair, B-2, is 1 bit. Hence, since the system as a whole produces 4 bits of information, while the sum of the information produced by its two parts independently from each other is about 2 bits, it emerges that the system as a whole is capable of producing integrated information.

Now, if consciousness is integrated information (as per the IIT), since the amount of integrated information produced by a system is quantifiable and can vary from system to system, it follows that consciousness can be quantified in turn. Furthermore, every conscious system will have its own amount of consciousness which is determined by the specific way its parts cooperate in producing integrated information. If this view proves feasible, this would be a remarkable advance in our understanding of consciousness. But what reasons are there to believe that consciousness is integrated information? One reason is that it is possible to interpret two aspects of the theory as corresponding to two important features of consciousness. First, the parts of an integrated system are not informationally independent from each other, meaning that the produced integrated information comes from the joint parts rather than from any of them taken in isolation, so that it is not possible to associate specific bits of integrated information with specific parts

of the system. This seems to capture the indivisibility of conscious experience, whose constituent elements cannot, in general, be selectively eliminated by tampering with this or that specific elementary unit of the realisers of the system (in the case of humans, specific neurons). Just like it is impossible to see a patch of colour without simultaneously perceiving a certain shape associated with it, it is not possible to take a relevant (group of) neuron(s) away without reducing the overall amount of information by a value higher than that generated by that individual component.

Furthermore, Tononi suggests that "complexes", i.e. those systems capable of producing integrated information which are not themselves entirely contained in any other system producing a higher amount of integrated information, are useful to account for the private character of conscious experience. For, the information processed by a complex as a whole cannot be taken to be processed also by another complex which fully contains the former. This is because, although complexes can partially intersect, not every part of a complex can be also a part of another, so that not all the information generated by a complex (neither as a whole nor as a non-integrated sum of parts) can be immediately accessed by another complex. Note that this also sits well with the idea that the information processed in cognition has to be accessed without mediation, from a 1PP. Complexes allow the formulation of a third-personal story about why this is the case, as a complex's information is intrinsically contained in it, and any information coming from outside a complex as a whole would not be the complex's own information, but, at best, something the complex's subsequent informational processing can carry information about.

To summarise, the IIT addresses some of the features of consciousness which have proved most difficult to account for, namely the indivisibility of conscious experience, and its inherently private character. This, together with some other aspects of it that will be presented in detail in the next subsection, makes the IIT a theory of consciousness capable of accommodating the 1PP condition.

3.2 The IIT as an axiomatic theory

Every scientific theory has to start from somewhere. In order to address their target phenomenon, empirical theories need to take their moves from some observations and

assumptions concerning the nature and properties of what will be described by the corresponding theory. For instance, cell theory needs to start from the observationally motivated assumption that cells exist, are the constituents of biological entities on Earth, and are responsible for an organism's life and reproduction. Upon such grounds, the theory can be developed and honed in such a way that it will ultimately be possible to formulate accurate explanations and predictions concerning its targets. However, the scope of a theory, i.e. the phenomena which it is about, must already be in place before a theory can be built. To be sure, this does not mean that anyone working towards the formulation of a theory must precisely know what sort of phenomenon is being studied before having some sketch of the relevant theory. Rather, this means that before theorizing on something one needs to broadly individuate the range of observable phenomena which will be theorized upon. For example, before coming up with a precise characterisation of what makes bodies fall to the ground, of how the fall works, and of why it works that way, one needs to first acknowledge that some bodies appear to naturally fall if left unconstrained.

Being no exception to this pattern, the IIT also needs to start from somewhere. Its peculiarity is, however, that it purports to be a theory of something observable only in a *sui generis* way. Differently from other natural phenomena, its target, consciousness, is directly observable only from a certain privileged perspective, the first-personal perspective, and it appears to possess features which are profoundly different from those shared by the vast majority of the rest of the objects populating our world. Given the peculiarity of the target of the IIT, its pre-theoretical individuation cannot be based on empirical considerations, in the objective, third-personal sense of "empirical". For this reason, the IIT is based on a set of phenomenally inspired axioms, from which the basic postulates of the theory are inferred and which I will present in turn. My presentation will be based on the more recent formulation by Tononi and Koch (2015) rather than on older versions (e.g. Tononi's (2012)).

Axiom 1 - Existence

The first axiom of the theory states that conscious experiences exist, and they exist intrinsically. That is to say, not only eliminativism about consciousness is ruled out in principle, but conscious experiences are taken to exist in and of their own, in an (external)

observer-independent way. Whether something is conscious or not does not depend on anyone establishing that this is the case; importantly, this includes the conscious subject as well, so that one can be conscious without being explicitly aware of this fact. In this sense, consciousness is not different from any other regular, naturalistically conceived magnitude. The existence and quality of conscious states are not something that can be arbitrarily stipulated. Accordingly, whether at some point a conscious state is occurring, as well as what its content is, are the sort of empirical facts an observer may be right or wrong about. One may attribute consciousness to something which in reality does not have any, and one may fail to acknowledge the presence of consciousness or misidentify its contents.

Axiom 2 - Composition

The second axiom of the IIT states that each experience is structured, that is, it is composed of many elements. This means that the content of each individual conscious experience contains a variety of pieces of information about a plurality of things. For instance, if my elbow is itching, I am feeling an itch, which in turn has a degree of intensity and is located somewhere around my elbow, which incidentally is implicitly given to me as mine. At the same time, I may have a visual experience of the laptop in front of me, which has its screen on, which is mostly white, and so on. In a nutshell, while each conscious experience happening at any particular time is a single experience, it contains information about a multiplicity of distinguishable states of affairs which one can potentially focus their attention on separately, although, as the fourth axiom will later highlight, not entirely independently from each other.

Axiom 3 - Information

According to the third axiom, conscious experiences are specific. That is to say, each conscious state has its own unique phenomenal properties, which make it distinguishable from any other conscious state. For instance, the mild feeling of relief one has in finding out that the bus one was meant to catch has not passed yet has a specific phenomenal character, which makes it identifiable as a mild feeling of relief, as opposed to a strong feeling, or to a different feeling like rage, or to a different sort of experience like a visual experience of a book (or of a blue book, or of a big book…).

This axiom has an important corollary, namely that, in conformity with Leibniz's Principle of the Identity of the Indiscernibles, if two conscious states experienced at different times have exactly the same phenomenal properties, then they are the same conscious state. To be sure, in the light of *Composition*, it rarely if ever is the case that one conscious state occurs more than once. It is indeed possible that one experiences a similar feeling of relief in discovering that one is still on time to catch the intended bus in more than one circumstance. But since experiences are composed of several different contents appearing with various degrees of intensity, it is highly unlikely that a conscious state as a whole will be identical to another conscious state as a whole occurring at a different time. One may feel relief on a sunny summer day while being soaked in sweat, or one may have an analogous feeling while shivering under the snow. Therefore, while some aspects of a conscious state may plausibly resemble aspects of other conscious states, it is unlikely (although not impossible) that two conscious states ever are phenomenally perfectly indiscernible, given the vast amount of information concerning a conscious system or its environment that consciousness can carry at any given moment.

Axiom 4 - Integration

The fourth axiom of the IIT is closely related to the second one. While the second axiom points at the fact that within a single conscious state there coexists a number of different pieces of information, the fourth axiom highlights instead how, despite the more or less variegated contents they may have, conscious states are integrated into unified wholes. In walking the aisles of a supermarket looking for cereals, one sees the great many colours of the packages surrounding them, feels the resistance opposed by the trolley and the cold air coming from the open fridges, hears the voices of the other customers, is thinking about the look of the sought after package, and so on. Nonetheless, all these different contents are not sharply separable from one another, as they all combine into the unified experiences one has at every moment. While one can tell the difference between the resistance opposed by the trolley from other parts of one's experience, such resistance is not perceived in isolation, in parallel with the other contents of one's conscious state. Rather, one has a single, unified experience, some of whose elements are more salient than others, some are less focused upon than others, but which remains unified and undivided.

It is important to notice that the second and the fourth axioms of the IIT are not in conflict with one another, as it may instead seem at first. The fact that conscious experiences are given as unified wholes does not prevent them from being composed of a number of discernible contents. This is important, because one may instead be tempted to point out that, if experiences really contain phenomenally distinguishable aspects, then they are not "really" unified. Rather, any conscious state one undergoes at any given time is a collection of independent, separate conscious elements which, like the members of a crowd of people, could exist individually. More precisely, one may be inclined to argue that conscious experiences as wholes are not unified, in the sense that they do not exist as legitimate wholes possessing properties in their own right, over and above the contributions of their parts, just like a body of water does not exist, properly speaking, as something more than the sum of hot and cold streams, of foamy areas on the surface, of more or less muddy waters in different points. However, this is not the way experiences appear to be conceived of through the axioms of the IIT. Conscious experiences as conceived by the IIT are irreducible to their constituent parts, just like a pointillist painting is irreducible to the sum of its tiny patches of colour, as the overall emerging shape depends on the particular relations that the parts entertain with each other.

Hence, just like in the case of a pointillist painting it is possible to discern, upon closer inspection, the numerous dots making up the overall painting without the painting thereby ceasing to exist as a unified whole for this reason, the second and fourth axioms of the IIT are not inherently incompatible. One can discern different phenomenally identifiable components within a single experience without this undermining the unity of the experience as a whole. Obviously, one may still want to challenge this view, which is taken by proponents of the IIT to be evident and thus left unargued for, but whether this claim is ultimately true or not does not matter for present purposes. All that matters here is that no inconsistency immediately arises from the second and the fourth axiom of the IIT.

Axiom 5 - Exclusion

The fifth, final axiom of the IIT states that consciousness is definite in content, that is, that each conscious state contains neither more nor less content than it appears to have. For instance, if one looks at a sliced pomegranate from a distance, one will have a visual

experience of a red slice of fruit, but it will not be possible to discern the single arils. According to the axiom of *Exclusion*, that visual experience does not contain the visual experience of any individual aril, as experiences do not contain anything more than they appear to have. Or, if one has a quick look at some geometrical shape with a relatively high number of sides (no need to appeal to chiliagons: probably even a decagon will suffice), one will see that geometrical shape, but the related conscious state will not contain any specific information about the number of sides of the shape.

One may object that in many cases our conscious states contain more information than they appear to have. For example, when one tries to remember what some movie's ending was, one may have a sudden glimpse, or "aha! moment", in which a particular frame of the ending is recalled. But that frame stands for, say, ten minutes of the movie, which one would then be able to recall in some detail if prompted to do so. In cases like that, one may be tempted to say that the iconic image suddenly popping up in one's mind has much more content than it prima facie appears to have. However, proponents of the IIT would probably reply that this is not the case. The relevant conscious state only contains a mental image with a distinctive associated feeling, but the information that is subsequently accessible after that brief epiphany is not, properly speaking, contained in the initial glimpse. The conscious state under discussion only serves as a gateway to the information concerning the whole ending of the movie. Despite being given to the subject as an iconic image standing for much more elaborate content, it only is a placeholder, and the lengthier content is not part of the relevant conscious experience.

Along similar lines, also the other part of the fifth axiom of the IIT may be questioned. After all, one may have the illusion that a given conscious state contains much more phenomenally given information than it really does. To see why this may be the case, one can think of the blind spot humans have in the right side of their visual field[33]. Without the vast majority of us being aware of this fact, and without any of us ordinarily noticing this anyway, there is a little blank area in the right lower section of our right retina which is automatically "filled in" with visual content by our brain, but in which we really are blind. Unless one pays special attention to it and actively tries to find that spot, for instance by closing the left eye and, while fixating a point with the right eye, slowly

[33] Another standard example would be peripheral vision, which contains a lot less information, in terms of details and colour, than most people imagine.

moving one's finger from left to right until its tip disappears, this fact is impossible to notice, and one will accordingly think that every part of one's visual field is filled with as much content as it appears to have. Therefore, it may be argued, at least some conscious states contain less information than they appear to have.

According to proponents of the IIT, this objection too would be misguided. It is true that, upon reflection, our conscious states may not be as truly informative as they seem to be, but this does not mean that their content is narrower than it appears. In the case of the blind spot, one does have the conscious experience of a continuous visual field, it's just that part of that visual field is fabricated. One does not see the world in that part of the visual field, so that what one would think being situated in it may or may not exist, but this does not mean that such illusory content is not part of one's experience: it just is something which corresponds to nothing in the world.

Summary

To summarise, the IIT is an empirical theory grounded on five axioms which play the role of bedrock truths (Bayne (2018)). They provide the conceptual foundations of the IIT and characterise conscious experiences as intrinsically existing (axiom 1) unified wholes (axiom 4), although composed of a multiplicity of phenomenal components (axiom 2) endowed with a specific content, which is neither narrower nor more extended than it phenomenally appears to be (axiom 5), and which makes them uniquely identifiable (axiom 3). Clearly, their self-evidential status can be and has been questioned. However, I will postpone the discussion of this and other criticisms to the fourth section. Before doing so, I will explain how the IIT is equipped to defend the claim about the 1PP from the somnambulism challenge.

3.3 The IIT and somnambulism

To recapitulate, the IIT posits a fundamental identity between consciousness and integrated information. A system is conscious to the extent that it is capable of generating integrated information, and, as a consequence, consciousness comes in degrees. This has interesting implications for the original objection based on somnambulism against the 1PP condition discussed earlier. The gist of the objection was that the 1PP cannot be

necessary for cognition if there is reason to believe that this leads to the further claim that there can be no unconscious cognitive activities. This is because the firmly established scientific consensus is that unconscious cognition can and does occur, one of the most notable examples of this being represented by somnambulists. Therefore, if the 1PP condition is to stand the objection, it needs to be coupled with a theory of consciousness able to fend it off. Hence the question: is the IIT up to the task?

Since the IIT identifies consciousness with integrated information, one of the IIT's predictions is that consciousness will fade if a previously integrated system ceases to be such. This prediction is corroborated by the growing understanding of NREM sleep in humans (and in other mammals). In fact, several empirical studies have in recent years highlighted how high-definition EEG and transcranial magnetic stimulation show that during NREM sleep we witness a breakdown of cortical effective connectivity (Tononi and Massimini (2008), Massimini et al. (2010), Pigorini et al. (2015)). This leads to an increased modularity of the brain (Boly et al. (2012)), where short-range, within-system neuronal network connectivity increases while, importantly, long-range, between-system connectivity decreases. As a consequence, during NREM sleep the capacity of the brain to integrate information on a large scale is substantially diminished. In the light of the IIT, this would explain why consciousness fades during NREM sleep, as the IIT establishes an identity between consciousness and integrated information.

It seems then that regular NREM sleep per se corroborates an important prediction of the IIT. Now, somnambulism is a disorder of arousal (Bassetti et al. (2000)) occurring precisely during NREM sleep. Therefore, the fact that it is possible to witness seemingly cognitively driven behaviours during NREM sleep, when consciousness is absent both as a matter of common sense and as per the IIT, would be all the more damning for the 1PP condition. But this consideration fails to factor in the recent consensus on the nature of sleep in general. In fact, while sleep typically is observable as a phenomenon controlled by global mechanisms, it is a highly localised phenomenon (D'Ambrosio et al. (2019)). Wakefulness and sleep, and in particular NREM sleep, are not mutually exclusive, and elements of each can co-occur in different portions of the brain (Krueger et al. (2019)), especially during somnambulism (Terzaghi et al. (2009), Terzaghi et al. (2012), Siclari and Tononi (2017), Zadra and Levitin (2022)). For instance, a study carried out by Castelnovo et al. (2016) shows that the EEG of somnambulists (and subjects suffering

from night terrors) significantly differs from the regular EEG associated with NREM sleep when sleepwalking episodes occur, locally displaying features associated with wakefulness. Similarly, Desjardins et al. (2017) report an increased long-range connectivity in somnambulists, which may enable at least a partial restoration of some degree of consciousness. This would corroborate Zadra's et al.'s (2013) observation that much more mentation than previously thought occurs during sleepwalking episodes.

To summarise, while they may constitute a prima facie objection to the claim that all cognition has to be conscious, the presumably cognitive activities performed by somnambulists are not at odds with such a claim in an obvious way, in the light of the current state of our knowledge of NREM sleep and sleepwalking. If one endorses the IIT as a theory of consciousness, one may have the means to defend the condition against this potential attack. There is a catch, though. The previous considerations only show that the IIT, combined with empirical studies on NREM sleep and somnambulism, has in principle the means to accommodate the cognitive status of the activities performed by somnambulists. However, they do not show that the IIT *does* succeed in doing so. For the IIT to properly accommodate somnambulistic cognition, a direct comparison between the quantities of integrated information associated with sleepwalking episodes, on the one hand, and with regular NREM sleep on the other should be carried out. This is where things get tricky, as exactly quantifying the amount of integrated information produced by real-world systems is a computationally intractable task, preventing the IIT's predictions from being properly empirically tested (Merker, Williford, and Rudrauf (2022a, 2022b)). In the next section, I will elaborate on this point. Together with the discussion of some other general issues associated with the IIT as a theory, this will ultimately lead me to conclude that, in its present form, the IIT is not suitable as a companion for the 1PP condition, notwithstanding its merits highlighted in this section.

4. Problems with the IIT

In the previous section, I have introduced the five axioms underpinning the IIT, and I have mentioned the fact that integrated information is not perfectly equivalent to Shannon information, as it purports to be an intrinsic quantity, possessed by integrated systems regardless of any external, more or less negotiable constraints. Much of the

discussion around the IIT has legitimately focused upon these two aspects of the IIT, namely its axiomatic guise and the "intrinsicness" of integrated information. The purpose of this section is to present some important objections against the IIT that have been moved with respect to these two aspects, as well as to the difficulty of empirically testing the theory. My conclusion will be that, as things currently are, even if the IIT would be in principle able to protect the 1PP claim, it is not sufficiently robust to actually do so.

4.1 Is the axiomatic guise appropriately developed?

The five axioms providing the conceptual backbone of the IIT are taken by Tononi and colleagues to be self-evident truths, meaning that they are supposed to capture phenomenally unquestionable features of consciousness. This is a rather bold claim, and it is not hard to see that each axiom can individually be challenged, starting from the very first one, the seemingly most solid of them. After all, one may be an eliminativist or illusionist about consciousness, thus denying the very existence of it. Regardless of whether this is a reasonable position to defend, there are, as a matter of fact, people upholding it (most notably, Dennett (1988, 2005), or Frankish (2016)), which is enough evidence to concede that the first axiom is not self-evident.

However, it is not crucial for the IIT that its axioms are indubitable truths. All that is needed is that they are factually correct, even if their correctness is not immediately evident, and even if their correctness is yet to be proved. For the purposes of the IIT proper, rather than for its foundations, it suffices that they are formulated in such a way that they can act as guiding principles from which it is possible to infer the actual postulates of the theory, that is, the more formal counterpart of the axioms based on which the theory can be systematically developed. These postulates spell out the ideas expressed by the axioms in causal terms and involving more technical notions such as the "cause-effect repertoire", i.e. the range of potential states a subset of the system, or the system as a whole, can enter in as a consequence of having previously been in some state. Thus, proceeding in the same order as the axioms, and again following Tononi and Koch (2015, p.7): the first postulate, corresponding to *Existence* (according to which consciousness is taken to exist intrinsically) states that a system existing intrinsically must be causally effective upon itself; the second postulate, derived from *Composition* (according to which

each experience is composed by a plurality of elements), states that collections of elements of the system must be causally effective upon the system as a whole. The third postulate, inspired by *Information* (the axiom stating that each conscious state is distinguishable from any other conscious state) states that a system must specify a particular cause-effect structure, that is, a specification of the cause-effect repertoires of all the subsets of the system. The fourth postulate is inferred from *Integration* (the axiom stating that each conscious state is a unified whole) and states that the cause-effect structure of the system must not be reducible to the causal repertoires of non-interdependent subsets of the system. Finally, the fifth postulate, inspired by *Exclusion* (the content of each conscious state is neither more nor less informative than it appears to be) states that the relevant cause-effect structure is the one that maximises integrated information.

According to proponents of the IIT, these postulates can be derived from the axioms of the theory. That is to say, leaving aside the correctness of the axioms in their own right, these postulates or some analogous formulation of them can be inferred from the axioms. The problem, however, as Bayne (2018) correctly observes, is that it is not clear what sort of inferential link exists between the axioms and the postulates of the IIT. Surely it is not a deduction, nor an induction; hence, it must be an abduction. But abductive inferences need to be motivated and should not be accepted independently from confrontation with competing abductive inferences, which in this case would consist in alternative sets of postulates. Thus, if the inferential link between axioms and postulates is an abductive one, much more work than has already been done is required.

This line of research needs to be developed along two axes. First, proponents of the IIT need to motivate the choice of adopting these particular postulates. For instance, what reasons are there to think that the first axiom, according to which a conscious system must exist intrinsically, is well-captured by the first postulate, according to which a system must be causally effective upon itself, in the sense of being able to influence its own later developments? Indeed, this seems a necessary condition for intrinsic existence (existence by the system's own lights), but perhaps it is not also sufficient. While one may accept that if *X* does not exert at least some causal influence over *Y*, then *X* is practically not existing for *Y*, it is debatable whether simply exerting a causal influence over something else ensures the acknowledgement of something's existence, so to say. Even though I am

not sure of the strength of the following counterexample, it seems sufficient that this causal influence is cancelled out by an equal but opposite one, or simply "masked" by some other more important causal influence, to make the first postulate look as if it is not enough to achieve what it is meant to, because there would be a causal influence as required, but it would be one whose actual outcome is null or neglected.

Perhaps there are no strong reasons for believing that the current postulates of the IIT are particularly good aside from the impossibility of formulating alternative postulates. If the axioms were such that no other set of postulates could be inferred from them, then, provided that there are good reasons for accepting them in the first place, they would act as sufficiently strong constraints on the theory for the IIT to be informative and for accepting its postulates. However, as Bayne (2018) has again rightly pointed out, this does not seem to be the case, at least given the current state of affairs. The architects of the IIT do not appear to have not really considered the possibility of formulating a different set of postulates, nor have they proved that it is not possible to do so. For this reason, further research is needed in this direction as well, for the IIT to be adequately robust as a theory. As it stands, the choice of its postulates appears to be underdetermined by its axioms, and this warrants at least some degree of scepticism with respect to its robustness as a theory of consciousness.

4.2 Is integrated information acceptable?

One of the most interesting features of the IIT is that it conceives of consciousness as a measurable natural phenomenon, not different from any other legitimate physical magnitude such as mass, charge or temperature. More specifically, a crucial feature of consciousness that contributes to its having this status is, on the basis of the IIT, its being able to exist independently from any external observer. Regardless of whether anyone is measuring it, if the first axiom of the IIT is correct, then consciousness still exists, and if the central identity posited by the IIT equating consciousness and integrated information is on the right track, it exists as the integrated information generated by a system.

In the previous subsection, I have pointed at two sorts of difficulties that arise for the IIT as a theory: its axioms may not be the self-evident truths its proponents take them to be, and the postulates of the IIT may not be the correct ones, because no alternative

sets of postulates have been explored. Here, I will focus instead on the central identity of the IIT, discussing an issue that arises in relation to it, and more specifically with respect to the notion of integrated information.

Arguably the most important problem with integrated information, which has been repeatedly noted and which the proponents of the IIT themselves acknowledge, is practical in nature. Namely, calculating the amount of integrated information generated by a system requires an unfeasible amount of computations, at least given the computational power available today and which will probably be available in the foreseeable future. Based on the fifth postulate of the IIT, the amount of integrated information produced by a system has to be calculated over the cause-effect structure that maximises it. That is to say, given a certain system, the quantity (and quality) of integrated information which is relevant for the purposes of the IIT is the one corresponding to only one of the cause-effect structures which the system may involve, namely the structure that produces the highest amount of integrated information. But, leaving aside unrealistically simple systems composed of just a small bunch of elements, there will be a very high number of such structures associated with any system. Not only that, but for each of them an exponentially growing amount of computations would be required, since one would need to consider how each state of the system organised in a given cause-effect structure constrains its cause-effect repertoire (i.e. how a given state determines which among the possible subsequent states of the system will occur). In the case of complex systems, composed of a huge number of highly interdependent elements such as the human brain and its neurons, the required computations are simply presently not feasible.

If this were the whole story, one could still hope for future technological advancements, which may one day enable us to compute the quantity of integrated information any system generates. But things are way more problematic than that, as it has been argued in a series of works (Barrett (2014, 2016); Barrett and Mediano (2019); Mediano, Seth and Barrett (2019)). Again, according to the fifth postulate of the IIT, the relevant cause-effect structure for the calculation of the integrated information of a system is the one that maximises integrated information. In order to do so, one needs to consider a definite system, which has clearly identifiable, discrete elements. In the case of the brain, one would probably be inclined to consider the brain as a system having individual neurons as its elements. The issue is that the choice of the fineness of grain to be adopted

in computing integrated information is underdetermined. In the case of the brain, even assuming that its physical boundaries are sharply defined, why should only cause-effect structures having neurons as elements be taken into consideration in calculating integrated information? As Barrett and Mediano have noted, the maximal amount of integrated information has to be calculated not only over all the possible cause-effect structures of a system at a certain level of description, that is, keeping a certain spatial graining fixed in the background. Rather, all the possible grainings have to be considered. Furthermore, such grainings are not to be understood in a purely spatial sense (e.g. molecules, aggregates of molecules, or brain areas, rather than neurons), but they also have to include various temporal grainings. Indeed, when one considers two subsequent states of a system, how far apart in time do they have to be for them to be considered "subsequent"? One millisecond? Forty milliseconds? One minute?

It should be clear at this point that the first difficulty concerning the currently available computational power rapidly degenerates into a truly intractable problem: computing the integrated information generated by a system would require an infinite, not just unfeasible, amount of computations. The reason why this is the case is that there is no standard way to choose some graining, both temporal and spatial, so as to be able to exclude the infinitely many others. And given that there is no standard graining to choose, picking one particular graining would be a completely arbitrary move, which would then jeopardize the IIT's most notable feature, namely its in-principle ability to provide a way of quantifying consciousness while at the same time conceiving of it as an observer-independent quantity. The choice of a particular graining over infinitely many others would amount to picking a framework in which consciousness exists in conformity with the limitations and preferences of an observer. If an ontology of discrete entities is presupposed, then consciousness/integrated information is either impossible to calculate, or it is not an intrinsic, observer-independent quantity.

Some attempts at overcoming the computational intractability of integrated information have been made. For instance, Barrett (2014) has suggested that a way to address this critical problem for the IIT may involve replacing the current ontology of the IIT, entirely composed of discrete entities, with an ontology which also has the benefit of being more in line with our current physical theories, namely one based on continuous fields rather than discrete objects. This is a viable research direction, but I will not try to

explore it here, as it is at present only a potentially viable direction where to find a solution to the problem, not a solution in itself. Or again, others (e.g. Barrett and Seth (2011), Mediano et al. (2019)) have explored alternative quantities acting as proxies for integrated information. However, as noted by Merker, Williford and Rudrauf (2021, p.3), finding a proxy for integrated information is a tricky task, as not all such measures are equivalent. In particular, they deliver conflicting results in terms of the hierarchical rankings among the complexes producing integrated information. This is problematic, as one of the tenets of the IIT is that only the complex producing the highest amount of integrated information is the one that actually supports the consciousness associated with the overall system.

Even though the issues I have discussed here may end up being solved, what matters for the present purposes is that the core identity of the IIT involves a quantity, integrated information, which is not well-defined for any real-world system. Given the current state of affairs, proponents of the IIT still have a long way to go before the IIT will be a mature, acceptable theory of consciousness. As a result, at present the IIT is not able to come to the rescue of the 1PP condition, despite having the potential, at least in principle, to do so.

Conclusion

It is now time to draw some conclusions. The open issues discussed in the previous section concerning the empirical testability of the IIT, as well as other theoretical problems having to do with its axiomatic systematization, prevent the IIT from saving the 1PP condition, with its implication that there cannot be unconscious cognition. Does this mean that we are back to square one, with all the apparent progress made towards a better understanding of cognition being wiped away? In a way yes, but not entirely, and at any rate not just yet. It is true that the claim that a 1PP is necessary for cognition has been undermined by the objections presented in this chapter, but that does not mean that all the insights gained along the journey have to be dismissed with it. First of all, the original observation from which the discussion developed in this chapter, namely that there seems to be a non-negligible similarity between the notion of 1PP as I have characterised it and the way the 1PP is treated in the phenomenological literature, does not necessarily lead to the conclusion that there cannot be unconscious cognition. That is a likely subsequent

claim, but one may simply reject the implication and save the 1PP condition. This is the reason why in the next chapter I will present a further argument against the 1PP condition. The outcome will be that we should dismiss that claim in the end, but this will not leave us with anything under our belt.

Chapter 5

Introduction

In this chapter, I will conclude my criticism of the 1PP condition I have been discussing extensively in chapters 3 and 4. This criticism will be structured in a way that it will not apply only to this specific condition on cognition, but also to the broader methodological family of which the 1PP condition is a product. This family is that of the anthropogenic approaches. i.e. those attempts at finding a mark of the cognitive (MOC) that take human cognition as the starting point for their inquiry. The outcome will be a recommendation not to pursue anthropogenic approaches in the search for a MOC and to pursue a biogenic strategy instead (i.e. one where the inquiry begins with living organisms, rather than just humans).

In the first part of this chapter, I will introduce the main players of my argument. Then, in section 2, I will explain what anthropogenic and biogenic approaches are, and how they differ from each other. In section 3, I will introduce the metaphysical principle that will play the most prominent role throughout the rest of the chapter, namely Alexander's Dictum, according to which something is real if and only if it has causal powers. In section 4 I will re-introduce the 1PP condition on the scene, and I will prepare the transition to the dilemmatic structure of the rest of the chapter. That is, I will argue that such a condition is best understood as the product of an anthropogenic approach, on the one hand, and as framed in nonreductionist, as opposed to reductionist, terms.

The dilemmatic structure of the rest of the chapter will revolve around the exploration of the consequences for a MOC of committing oneself alternatively to nonreductionism or reductionism. On the one hand, going nonreductionist would require, in the light of Alexander's Dictum, the attribution of distinct causal powers to the higher-level properties claimed to be necessary for cognition. But in section 5 I will argue that the familiar (causal) exclusion problem would hit nonreductionist proposals hard, and I will reply to a way of addressing this problem which has recently gained significant traction, consisting in the adoption of the interventionist account of causation. Furthermore, in section 6 I will consider, and dismiss, a possible way to disarm the

epiphenomenalism imposed on nonreductionist MOC's by the exclusion problem. This would consist in the adoption of a minimal form of naturalism that may allegedly undercut Alexander's Dictum, which, in turn, is what arguably makes epiphenomenalism problematic. However, I will show that this minimal form of naturalism cannot be paired with the 1PP condition on cognition, and that, more generally, if Alexander's Dictum were rejected, a nonreductionist MOC elaborated in such a context would not be acceptable.

Thus, the first part of the dilemma will result in the dismissal of nonreductionism. This will make the 1PP condition unacceptable, but it will not speak against the anthropogenic approaches more generally. It is in section 7 that the second, reductionist horn of the dilemma will be discussed. I will argue that, typically, the products of anthropogenic approaches do not sit well with the combination of reductionism with Alexander's Dictum. This is because the higher-level conditions on cognition advanced by anthropogenic MOC's generally lack a naturalistic characterisation. This means that it is not possible to point at the causal powers of the potential realisers of the relevant properties to substantiate anthropogenic claims. Thus, considering that such higher-level properties do not have clear causal powers associated with them, and provided that Alexander's Dictum is a solid principle, also reductionism is not viable for anthropogenic MOC's. Therefore, since anthropogenic approaches can be paired neither with nonreductionism nor with reductionism, they should not be pursued.

Finally, in section 8, I will argue that the situation is different for biogenic approaches. They too should not be conceived of in nonreductionist terms, as my criticism of nonreductionism applies across the board, not just in the context of anthropogenic approaches. But reductionism does not lead to any significant difficulty for biogenic approaches, even if Alexander's Principle is kept fixed in the conceptual background. Thus, the second horn of the dilemma, namely the one exploring reductionism, highlights how biogenic approaches are viable, differently from their anthropogenic counterparts. The search for a MOC should therefore be carried out in biogenic terms.

1. Anthropogenic vs. biogenic approaches, and the notion of "levels"

If one is in the business of identifying one or more individually necessary and/or jointly sufficient conditions for the occurrence of cognition, there are, broadly speaking two methodological strategies available. Their inquiry can take its moves from already acknowledged exemplars of cognitive systems, and then proceed to establish which features of those cognitive systems' behaviours and mechanisms are inherently relevant to their status as cognitive systems. Alternatively, one can address a broader domain of entities, such as biological organisms, some of whose members display cognitive phenomena, and then work one's way towards a minimal set of properties and features which appear to be sufficient and necessary for those members to manifest cognition. The first of the two methodologies is instantiated, within Lyon's (2006) taxonomy, by the family of the anthropogenic approaches, and it proceeds in a top-down manner. The latter instead, which moves along a bottom-up direction, is exemplified by the family of the biogenic approaches.

Besides being developed in a top-down or bottom-up fashion, accounts of cognition that belong to either of these methodological families do not share a precisely specified set of features, in the sense that, aside from the common general approach (biogenic approaches all begin their inquiry from living organisms in general, while anthropogenic approaches start from human cognisers), any two accounts instantiating the same approach may differ greatly from one another. As a result, being an anthropogenic (or biogenic) MOC, as opposed to a biogenic (or anthropogenic) one, ought to be conceived more in terms of fitting within a cluster of theories rather than in terms of membership to a well-defined class. For instance, based on the most common takes on cognition emerging from the anthropogenic camp as outlined by Lyon, it is quite frequent for anthropogenic MOC's to claim in one way or another that cognition involves intentionality (e.g. Adams and Aizawa 2001, 2008), or that cognition has to do with information processing (e.g. Rowlands 2009, 2010). Similarly, biogenic accounts of cognition typically claim that cognition plays an important role in regulating the causal interactions of the cognitive system with its environment in a way that promotes the system's adaptability and, consequently, survival (e.g. Keijzer 2021; but also Kiverstein and Sims 2021; or Van Duijn et al. 2006), but they differ from each other with respect to how this is exactly cashed out.

It is reasonable to trace the origin of the differences among the claims associated with the anthropogenic and biogenic families back to the overall methodological axes along which the various respective accounts are developed. If one begins their inquiry taking humans as paradigmatic cognitive systems, it will be more likely that one will define cognition in terms of properties or phenomena which are closely tied to human cognition than if one were to start from other, simpler organisms, such as bacteria or moulds, whose cognitive status is uncertain, or whose associated putatively cognitive phenomena differ greatly from those displayed by humans. The latter methodological approach, characteristic of the biogenic family, will instead tend to deliver results in terms of the intelligent ways in which the biological mechanisms constituting some organism promote the clearly biological needs and goals of that organism. For instance, anthropogenic approaches display a stronger tendency than biogenic approaches to express their claims about cognition in specifically psychological terms; on the other hand, biogenic approaches tend to express their claims in a more strictly biological conceptual framework.

As a result, anthropogenic accounts of cognition generally propose one or more individually necessary and/or jointly sufficient conditions for the occurrence of cognition that can be regarded as being "higher-level" than those proposed by biogenic accounts. The notion of 'level' employed here is not the one proposed by Oppenheim and Putnam (1958), where levels correspond to specific disciplines (the most fundamental one is physics, the level immediately above is that of chemistry, then biology, then psychology…). It may well happen that higher-level entities fall within the scope of some discipline's inquiry while some lower-level entities fall within the scope of another, but being studied by some discipline rather than another is not, in the present context, what places an entity or process at a higher level than another entity or process. Rather, the relevant notion of "level" is more similar to Wimsatt's (1994) one, which, as reported by Bechtel (2008, p.145), allows for a hierarchical ordering of the levels based on both size (smaller things are at lower levels than bigger things) and mereology (the parts of a whole are at a lower level than that of the whole). Hence, saying that the properties individuated as candidates for a MOC are captured in higher-level terms means that they are properties possessed by entities which can be further decomposed into smaller parts, each involved in their own activities. To give an illustrative example: the level at which a flower's

property of having a certain smell is placed is higher than the level to which the molecules responsible for that smell belong. This is so not in virtue of the sort of property "smelling in such and such a way" is, but because the flower's having that smell is a state of affairs that can be broken down into more fine-grained states of affairs, for example someone's odour receptors being stimulated by such and such odorous molecules.

As anticipated, I will be concerned with the examination of some theoretical difficulties that anthropogenic accounts of cognition face in virtue of their advancing claims about cognition involving higher-level properties and states of affairs. Such difficulties arise from considerations about the causal powers of the properties or processes claimed to be necessary for cognition by the individual MOC's. Therefore, before being able to discuss such difficulties, I will introduce the theoretical principle from which the tensions affecting the members of the family of anthropogenic approaches originate.

2. Marks of the cognitive and Alexander's Dictum

To avoid including trivial or irrelevant properties in a MOC, such as those simply correlated to cognition without having any particular significance for it, the properties in question should plausibly be instantiated by cognitive systems in virtue of some contribution they make towards the occurrence of cognition. It is natural to think of such contribution as involving causality to a certain extent. This is because some property is necessary for cognition not because it causes cognition to occur, but because it is constitutive of what counts as cognition. Given the above characterisation of levels, if a property is constitutive of cognition, then either it has to be the sort of property brought about by certain causally driven activities carried out by some lower-level mechanism, or it is the property itself that plays a role within the causal concatenations which cognition consists in. Hence, some property is necessary for cognition in that it is constitutive of it in a sense that involves causality to a certain extent. For instance, if one were to argue that intentionality is necessary for cognition, one may want to say that this is because states endowed with intentional content are what makes it the case that supposedly cognitive systems engage in cognitive activities, which in turn consist in operations

performed by whatever unfolding of states of affairs realises such content. Hence, it seems a reasonable provisional stipulation that:

(1) If a property is claimed to be necessary for cognition as it is constitutive of cognition, that property must be related to causal processes, either in its own right or because the corresponding lower-level states of affairs are engaged in causal processes.

However, the plausibility of (1) does not mainly derive from the previous considerations. Rather, the primary reason why it is a plausible requirement for a MOC is that it is a particular case of the more general principle known as Alexander's Dictum (see Kallestrup 2006, p.468):

(Dictum) For a property to be real, it must have causal powers.

(Dictum) is generally taken to be a reasonable principle, because, provided that one is on board with a broadly naturalistic philosophical approach (perhaps even with physicalism), one will be inclined to agree that, were an entity (object or property) causally inert, it would not be relevant to anything in the world, thus making its very existence dubious. For how could it be possibly detected, and what reasons would there be to suppose that it exists?

If one accepts (Dictum), one has also to accept that, whatever property is characteristic of some natural phenomenon, that property must be real in (Dictum)'s sense. This is because, otherwise, that property would not be capable of substantially characterising the phenomenon in question, as there would not be reason to take something that does not possess causal powers (hence whose very claim to existence is unfounded) as constitutive of a natural phenomenon. In the present context, the natural phenomenon in question is cognition, which, accordingly, plausibly consists in the unfolding of causally driven processes, whatever such processes may be. Hence, given that cognition is a natural phenomenon, a stronger version of the above principle follows:

(1*) If a property is claimed to be constitutively necessary for cognition, that property must possess causal powers.

In other words, by (Dictum), it is not just the lower-level mechanisms corresponding to the higher-level properties mentioned in a MOC that must be capable of causally acting and being acted upon; it is those very higher-level properties that must be capable of that. Otherwise, they would not be real, and cognition, which is instead here assumed to exist, could not be constituted by their occurrence.

Now, my ultimate target are anthropogenic (higher-level) MOC's, that is, MOC's elaborated taking human cognition as their starting point, and which feature properties possessed by entities and states of affairs that can be further decomposed into smaller parts performing different functions. As it happens, one can adopt a reductionist or a nonreductionist approach to this layered model, and in particular with respect to the causal powers of the higher-level elements mentioned in a MOC. In the next section, I will turn to the 1PP condition that I have explored in the past two chapters, and I will show that it can be legitimately considered an instance of an anthropogenic and nonreductionist condition for cognition.

3. The 1PP requirement as an anthropogenic, nonreductionist condition

In the previous chapters, I have been discussing the possibility that a necessary component of cognitive processes is that they involve a first-person perspective. Differently put, every cognitive process must have a 1PP associated with it, and, conversely, if there is no 1PP associated with some process, then that process cannot be cognitive in nature. In this section, I will show how such a claim can be considered the product of an anthropogenic approach. Not only that, but in the light of the particular way in which I have suggested that we should understand this claim, it should be interpreted in nonreductionist terms. Establishing these two points is the first step that will ultimately lead to the final reason why the 1PP condition should not be endorsed.

3.1 The 1PP condition is an anthropogenic claim

The claim that the 1PP is a necessary component of cognition was motivated by two main sorts of considerations. First, there are epistemic reasons. Given our pretheoretical understanding of the general characteristics of the target phenomenon, it

does not seem possible to conceive of cognition if not as of something that has a first-personal dimension associated with it. All the paradigmatic cognitive phenomena and processes appear to involve the processing of information in such a way that that information is accessed by a cognitive subject immediately and in a private way, so that the subject of such cognitive phenomena has a privileged kind of access to that information. Not only that, but since it also seems that cognitive subjects are not something distinct from the collection of the cognitive phenomena ascribed to them, this privileged kind of access appears to be explained by the fact that cognitive subjects are identifiable with the collection of the relevant cognitive phenomena itself. Therefore, the 1PP appears to be an ineliminable component of cognition, insofar as one understands the 1PP associated with a cognitive process as the "relation"[34] that exists between the process and its subject, which allows the information involved in such process to be made available to the subject in an immediate and private way. This is because cognitive processes without a 1PP associated with them would be subjectless, and that would clash with the intuitive understanding of what cognition is like.

On the other hand, I have also examined more metaphysical reasons for endorsing the requirement. That is to say, there is reason to think that cognition could not occur if not with a 1PP associated with it, even if one were to put aside the issues related to our understanding of cognition that would arise if we allowed for the possibility of the existence of cognition without a 1PP associated with it. This is because, were we to take the 1PP out of the equation, the individuation of the boundaries of cognitive systems would be underdetermined, so that drawing them in one way rather than another would plausibly be a matter of explanatory interests rather than of finding out about how the world really is. In this sense, an element of arbitrariness would enter the picture, and this is arguably not admissible if we take the existence of cognition as a natural phenomenon in a realist, ontologically serious way.

These two sorts of considerations that motivated the 1PP condition on cognition are the product of an anthropogenic approach. The first line of argument is rather straightforwardly so. Anthropogenic approaches take human beings as their starting point and they then proceed to individuate the constitutive features of human cognition. This is

[34] I am here using the term "relation" in somewhat of a reflexive sense, as cognitive processes are not distinct from their subject.

exactly what the first line of argument does: starting from the conception of cognition that one can obtain through introspection (i.e. through the observation of human cognitive activities "from the inside"), a claim about the nature of general cognition is put forward. The second argumentative strategy still qualifies as anthropogenic, but in a slightly less evident way. The anthropogenic component does not emerge clearly throughout, but primarily in the last step of what amounts to an inference to the best explanation. While the argument to the effect that the determination of the boundaries of a cognitive system should not be just a matter of explanatory interests is not peculiar to anthropogenic approaches, it is the proposed solution that makes the overall strategy an anthropogenic one. In fact, the boundaries of cognitive systems should plausibly correspond to what contributes to the occurrence of a 1PP, which is a notion that can be extrapolated from anthropogenically sourced considerations. In this sense, then, also the second route to the 1PP criterion can be considered an anthropogenic one.

3.2 The 1PP condition is a nonreductionist claim

There is also reason to think that the 1PP condition is to be thought of as a nonreductionist claim, not just as an anthropogenic one. This is a way of thinking of this claim that I have already discussed in previous chapters, where I have pointed out that the 1PP does not appear to be reducible to the 3PP. However, it is important to go a bit more in-depth here about what this amounts to.

There is a sense in which the irreducibility of the 1PP to the 3PP has to do with the notions of 1PP and 3PP from a purely conceptual point of view. To say that the 1PP is not reducible to the 3PP is to say that the properties that characterise the 1PP cannot be fully accounted for in terms of the properties associated with the 3PP. In particular, I have characterised the 1PP as a non-thematizing perspective, i.e. as a perspective that a cogniser has over their own cognitive states without this resulting in a distinction between the cogniser and the cognitive states themselves, while I have characterised the 3PP as a thematizing perspective, i.e. as a perspective that entails the distinction between the subject who entertains it and the phenomena over which it is entertained. Given this picture, it is clear that the 1PP cannot be reduced to the 3PP, because they are characterised as inherently different.

However, the fact that the concept of 1PP cannot be reduced to the concept of 3PP does not automatically mean that the states of affairs which have a 1PP associated with them are not reducible to states of affairs on which a 3PP is instead had. If that were the case, that would be tantamount to claiming that it is in principle impossible for cognitive phenomena to be reducible to non-cognitive phenomena, even just to the material realisers of those cognitive phenomena themselves, which is far from obvious. Were it to be the case, then it would be impossible, for the cognitive state I am in while having some visual experience, to be reducible to the physical substrate (whatever that may exactly be) of that cognitive state. While one may maintain that, as a matter of fact, there is reason to think that such a reduction cannot be operated, it is too strong to say that this is unavoidably so, as a matter of conceptual necessity based on the notions of 1PP and 3PP. Indeed, it may well be the case that cognitive states (which have a 1PP associated with them, according to the claim under discussion) are reducible to non-cognitive states (which only can have a 3PP associated with them). This is not impossible in principle, because the reduction of something which has a 1PP associated with it does not mean that the 1PP is to be reduced to the 3PP *per se*, but to states of affairs which can only have a 3PP associated with them. For example, if one were to try and reduce a cognitive state with a first-personal phenomenal component associated with it (say, a visual experience) to some neurophysiological state of affairs, that would not consist in reducing the phenomenal component of that cognitive state to the notion of "non-phenomenality" as such. Rather it would consist in reducing it to certain states of affairs as not associated with a first-personal phenomenal component.

In summary, the irreducibility of the 1PP to the 3PP does not entail that no state or process with a 1PP associated with it is irreducible to some other state or process without a 1PP associated with it. With that being said, I still think that the 1PP condition ought to be interpreted in a nonreductionist way. Based on the characterisation of the 1PP I have offered earlier, it does seem that the properties possessed by cognitive states in virtue of having a 1PP associated with them are taken to be irreducible to properties entirely characterisable in third-personal terms. In fact, under a standard reading of reduction that consists in establishing an identity between the reduced entity or property and the reducing entity or property, insofar as the existence of a 1PP is believed to be necessary for something to count as cognitive, the existence of a 1PP must not be identical with

states of affairs that instead do not involve the existence of a 1PP. For otherwise the existence of the 1PP would not make any specific contribution towards the instantiation and unfolding of cognitive processes, thus undermining the claim about its indispensability for the occurrence of cognition. This point can be phrased explicitly in terms of (Dictum). As mentioned above, if something is necessary for cognition, then it must make some specific causal contribution towards the occurrence of cognition. This is for two reasons. First, because, in the light of (Dictum), that property must be causally relevant in one way or another; if it were not, then it would not be a real property, and it could not therefore be constitutive of cognition, which is a real, natural phenomenon. Second, because if the causal manifestations of that property are indistinguishable from those of some other property, then there would be no reason to take those two properties to be distinct, and, as a consequence, there would not be any reason to claim that the former is specifically necessary for cognition. As a consequence, if the presence of a 1PP is claimed to be necessary for cognition, in the sense of being constitutive of it, it follows that the instantiation of a 1PP is not identical with some purely third-personal property or state of affairs. This means that, if the standard reading of reduction in terms of identity is acceptable, the instantiation of a 1PP is not reducible to any purely third-personal property or state of affairs.

In summary, it seems that the 1PP condition is to be understood in nonreductionist terms. Not only the 1PP is not reducible to the 3PP, but also the property "being associated with a 1PP" is not reducible to any property not involving the instantiation of a 1PP. Of course, one may not be persuaded by my line of argument, so that one may believe that requiring the existence of a 1PP can be also interpreted in reductionist terms. As it happens, if that objection were correct, the outcome of my argument against the anthropogenic approach to the research for a MOC would not be affected, although for different reasons. I will return to this point in section 7. For now, I will focus on the issues that the 1PP condition faces if interpreted as an anthropogenic, nonreductionist proposal on cognition.

4. The causal exclusion problem

It seems then that the 1PP condition is an anthropogenic proposal, and in particular one to be interpreted in nonreductionist terms. The latter qualification paves the way for a well-known argument against nonreductionism from the philosophy of mind, namely the causal exclusion argument, most famously associated with the name of Jaegwon Kim (e.g. 1998, 2007). In this section, I start by briefly outlining the argument itself, highlighting how it applies to the 1PP condition on cognition and how it appears to be especially problematic in the light of Alexander's Dictum. Then, I will discuss a possible way to fend off this attack, but I will show that this attempt does not succeed. The interest in discussing such an issue related to the 1PP condition, however, ultimately lies in the fact that it concerns such a condition not because of the specific claim that it makes about the nature of cognition. Rather, it lies in the fact that similar considerations can be made, *mutatis mutandis*, to any nonreductionist, anthropogenic MOC, as I will argue in the next section.

4.1 The causal exclusion argument

The gist of the causal exclusion argument in its original guise is that nonreductive physicalists are forced to accept the epiphenomenality of mental properties. However, as Kim (1997) has argued, the argument can be generalized: the threat of epiphenomenalism concerns not just mental properties, but all sorts of higher-level properties, if one embraces nonreductionism.

The causal exclusion argument is built from three core elements (Bennett (2003)). First, a principle of causal closure is established, whereby each physical event which has a cause has a sufficient physical cause (*Closure*). Second, some sort of determination relation between lower-level and higher-level phenomena (events or properties) is established (*Determination*). In the case of the 1PP condition, this amounts to two claims: first, the occurrence of a 1PP can only happen if certain physical states of affairs are in place; second, whenever one such physical state of affairs occurs, also the 1PP occurs. Finally, the last premise of the argument is one to the effect that causal overdetermination cannot be the norm when a causal process involving the higher-level property in question takes place (*Exclusion*). That is to say, in the case of the 1PP condition, one bars the

possibility that, whenever the 1PP is causally sufficient for the occurrence of some event, the lower-level state of affairs responsible for the occurrence of the 1PP is also causally sufficient for that event.

These three elements (C*losure*, *Determination* and *Exclusion*) jointly lead to the conclusion that higher-level properties or events, and the 1PP in particular, are epiphenomenal. To see how this works, consider a scenario in which a cognitive phenomenon *C* causes the occurrence of some event *E*. For simplicity, we can suppose *E* to be a physical, non-cognitive event. Since the 1PP is claimed to be necessarily constitutive of cognition, because of Alexander's Dictum there must be at least some cases in which cognitive phenomena cause the occurrence of other phenomena in virtue of having a 1PP associated with them. Let us then suppose that this is one of these cases: the 1PP associated with *C* is, in this particular case, the sufficient cause of *E*. Now, the instantiation of the 1PP associated with *C* (call it "*1PP*") is a higher-level phenomenon, in the sense that *C*, as a whole, is the kind of thing that can be broken down into a series of smaller mechanisms. Because of *Determination*, some of these mechanisms, *P*, are responsible for the occurrence of the 1PP associated with *C* (i.e. of *1PP*). This means, on the one hand, that the instantiation of *1PP* cannot happen without the occurrence of *P* (or of some other similar mechanism *P**), and on the other hand, that the occurrence of *P* necessitates the occurrence of *1PP*, so that also *P* cannot occur in the absence of *1PP*. Furthermore, since we are operating in a nonreductionist framework, *1PP* is not identical to *P*.

The next step consists in playing out *Closure*. Since *E* is a caused physical event, it must have a sufficient physical cause. Hence, since *1PP* is assumed to be a sufficient cause of *E*, the most plausible sufficient physical cause of *E* is *P*, namely the lower-level mechanisms responsible for *1PP*'s occurrence. But this would mean that *E* has two sufficient causes, *1PP* and *P*. Therefore, since this is assumed to be a typical case in which the 1PP manifests its causal powers, whenever *1PP* does so, *E* is causally overdetermined by *1PP* and *P*. This, in the light of *Exclusion*, is not acceptable: either *1PP* or *P* must be ruled out as a sufficient cause of *E*. But since, by *Closure*, there must be some sufficient physical cause of *E*, it is *1PP*, rather than *P*, that must be dismissed as a cause of *E*. As a consequence, the initial assumption that *1PP* is causally sufficient for *E* has to be rejected. In other words, *1PP* would be an epiphenomenon. This, at last, is especially troublesome

in the light of Alexander's Dictum, because epiphenomena are not real and cannot be constitutive of natural phenomena. Therefore, the claim that the 1PP is necessary for cognition because it is constitutive of it must be dismissed.

Before moving forward, an important concern needs to be addressed. A crucial presupposition for the argument to work is that the 1PP is not, as a higher-level property or state of affairs, physical in nature. One may thus wonder what motivates this assumption. In response, I think that it is important to keep in mind that, while physically realised, the 1PP as a higher-level state of affairs is here assumed to be non-identical with its physical realises. As a consequence, its physical or non-physical status is independent of that of its realisers. But, while this is no obstacle in the case of many other properties or states of affairs, in the case of the 1PP specifically (although not exclusively) this independence gets in the way of a physicalist conception. For, what would warrant the claim that the 1PP is physical in nature, if one cannot rely on its being physically realised? If one is to disregard the physical nature of its realisers, in that it has no bearing over the nature of the 1PP, one would need positive reasons to take the 1PP as such, i.e. as a higher-level state of affairs, to be physical. But those reasons are hard to find: what physical magnitudes associated with the 1PP, not just with its realisers, could one point at, in order to motivate their conception of the 1PP as something physical? I think that the answer is "none". The 1PP as a higher-level property is just not the sort of thing that falls within the domain of physics.

Let us go back to the exclusion argument. The literature on the topic is vast, and I will not attempt here to review it. Some have criticised the reliance on the existence of a "bottom level of reality" ensuring that causal powers are ultimately grounded (Block (2003)), but that is not a strong line of argument in the present context, given the characterisation of the layered model I am endorsing here. Some have taken issue with *Closure* (e.g. Lowe (2000)), but that is not too popular a strategy, because familiar considerations concerning the conservation of energy can be made on the one hand, and because abandoning *Closure* would likely require abandoning some intuitive form of physicalism[35]. Much more frequently, philosophers have focused their attention on

[35] Note that by "physicalism" here I do not intend the claim that everything in the world is physical in nature, for that would be immediately at odds with what I have just pointed out about the 1PP. Rather, I take this intuitive form of physicalism to maintain that everything in the world ultimately depends on physical happenings, in one way or another.

Exclusion, and in particular on the notion of causal overdetermination (e.g. Sider (2003); Bennett (2003)). In what follows, I will discuss solely what appears nowadays to be the most popular way to disarm the exclusion problem, consisting in the adoption of what is emerging as the default account of causation: Woodward's interventionist account. I will show how the interesting way of avoiding the epiphenomenalist pitfall based on interventionism does not succeed in securing causal powers to higher-level phenomena.

4.2 Can interventionism save the day?

In recent years there has been a number of attempts (e.g. Woodward (2008); List and Menzies (2009); Menzies and List (2010); Zhong (2014)) to disarm the exclusion problem by appealing to a specific account of causation, namely Woodward's (2003) interventionist account of causation. The core idea behind this account of causation is that, very roughly, *X* is the cause of *Y* if and only if, by causing the presence or absence of *X* while holding the background conditions fixed, one can make *Y* analogously present or absent. Slightly more precisely, *X* is the cause of *Y* if and only if both of the following conditions are met: (1) whenever some causal intervention *I* makes *X* present (*ceteris paribus*), also *Y* is present; (2) whenever some causal intervention *I* makes *X* absent (*ceteris paribus*), also *Y* is absent.

As Zhong (2014) has brilliantly shown, adopting this account of causation has two interesting effects. First, it makes both downward and upward causation impossible. Second, it allows for the causal autonomy of the higher levels, thus dodging the epiphenomenalist charge. Let us see how this works.

Suppose that some physical event, P_1, causes some other physical event P_2, and let us suppose that the occurrence of some higher-level event H_2 is determined by the occurrence of P_2. By the first clause of the interventionist account of causation, if some causal intervention makes P_1 present, also P_2 will be present. But by *Determination*, the presence of P_2 leads to the presence of H_2. Therefore, any intervention that makes P_1 present also makes H_2 present: the first of the two conditions that P_1 needs to meet to qualify as the cause of H_2 is met. However, the second clause is not met, if we allow for the multiple realisability of higher-level events. This is because an intervention that makes P_1 absent will also make P_2 absent, but the absence of P_2 does not entail the

absence of H_2. Perhaps the occurrence of H_2 is determined by the occurrence of an alternative realiser, say, P_3. It is thus possible that H_2 occurs in scenarios where some intervention made P_1 absent, so that P_1 does not qualify as the cause of H_2 in virtue of not meeting both requirements. Consequently, the interventionist account of causation is incompatible with upward causation, if multiple realisation is admitted.

Similar considerations can be made with respect to downward causation. Suppose that some higher-level event, H_1, causes some other higher-level event, H_2. By *Determination*, H_2 is determined in its occurrence by some physical event, P_2. Now, does H_1 cause P_2? If some intervention that makes H_1 absent is made, since H_1 is the cause of H_2, it would follow that H_2 would be absent too as a result. Furthermore, if H_2 is absent, also P_2 would have to be absent, because P_2 would otherwise determine the occurrence of H_2. Therefore, an intervention that makes H_1 absent makes also P_2 absent: the second condition for H_1's being the cause of P_2 is met. However, the first condition is not met. If an intervention that makes H_1 present were to occur, that would make also H_2 present. But the presence of H_2 does not entail the presence of P_2, if we admit that H_2 can be determined by some physical event $P_3 \neq P_2$ (i.e., if we allow for multiple realisation). Hence, the interventionist account of causation is incompatible with downward causation.

The interesting final bit of Zhong's argument is to the effect that, despite barring both upward and downward causation, the interventionist account of causation allows for intra-level, higher-level causation. Imagine a scenario where P_1, which determines H_1, is the cause of P_2, which determines H_2. Can H_1 be the cause of H_2? Let us run the test. Suppose that H_1 is the cause of H_2, and suppose that some intervention makes H_1 present. Since H_1 is assumed to be the cause of H_2, also H_2 will be present. At the same time, if H_1 is determined by P_1, then also P_2 will be present (because the former is a cause of the latter). If, instead, H_1 is determined by some other physical event *not-*P_1, then P_2 will be absent; but H_2 is present, and it must be determined by some physical event *not-*P_2. Either way, P_1 is confirmed as the cause of P_2, because its presence leads to the presence of P_2, and its absence leads to the absence of P_2. No contradiction is reached.

Finally, suppose again that H_1 is the cause of H_2, but suppose this time that some intervention makes H_1 absent. Since H_1 is absent, both H_2 and P_1 will be absent: H_2 is absent because H_1 is its cause, and based on the interventionist account an intervention making the cause absent will make the effect absent; P_1 is absent because its presence

would make H_1 present, and this is not the case by assumption. Not only that, but in this scenario also P_2 would be absent, for two reasons. First, because its cause P_1 is absent and, again, this entails P_2's absence, based on the interventionist account. Second, because if P_2 were present, then also H_2 would be present, and we know that this is not the case. Therefore, no contradiction is reached in this case too.

To summarise, Zhong (2014) has cleverly shown that if one adopts Woodward's interventionist account of causation (and multiple realisation is admitted), the epiphenomenalist conclusion does not necessarily follow from the exclusion argument. This means that one *can* be a nonreductionist and yet claim that the higher-level properties or events one is nonreductionist about are causally effective. Epiphenomenalism is not inescapable.

Of course, one may block Zhong's argument by either rejecting the interventionist account of causation or by denying the multiple realisability of higher-level properties or events. But neither option is palatable. Both moves would go against widespread consensus, and alternative accounts of both realisation and causation would be needed. Therefore, I will grant both components of Zhong's argument. Nonetheless, while undeniably representing significant progress in the debate over the exclusion problem, I do not think that Zhong's interventionist argument succeeds in turning the exclusion argument on its head. It is true that it makes the causal autonomy of higher levels possible, but this possibility does not entail that higher-level entities (or properties, or states of affairs) are, in fact, causally effective. Indeed, it seems to me that, in the interventionist argument's own terms, we do not have reason to believe that there actually are higher-level causal processes autonomous from the physical level processes they are determined by. Let me elaborate.

Consider some physical event P_2. If P_2 has a cause, by *Closure* it must have a physical cause, P_1. Now, suppose that P_1 determines the higher-level event H_1, and P_2 determines the higher-level event H_2. Because of the interventionist argument presented above, neither H_1 is the cause of P_2, nor P_1 is the cause of H_2. So, if H_1 is to avoid being considered an epiphenomenon, it has to be the cause of H_2. That is to say, the following two conditions must hold: when an intervention makes H_1 present, H_2 must be present too; when an intervention makes H_1 absent, H_2 must be absent too. Let then there be an intervention which makes H_1 absent. Since H_1 is absent, that intervention will also make

P_1 absent (because the presence of P_1 would determine the presence of H_1). Not only that, but the absence of P_1 entails the absence of P_2. Now, does the absence of P_2 entail the absence of H_2? Since I am granting that higher-level events can be multiply realised, no. It is indeed possible that H_2 will be absent, but it may as well be the case that it is present, because, say, it supervenes on some physical state of affairs P^* which occurs as the result of P_2's absence. Therefore, it is possible for an intervention to make H_1 absent without thereby leading to the absence of H_2.

Similar considerations can be made if we consider an intervention that makes H_1 present. The presence of H_1 does not guarantee the presence of P_1, because H_1 could be realised by some other physical state of affairs P'. If that is the case, then P_2 will be absent. But, again, the absence of P_2 may or may not lead to the absence of H_2, depending on whether some physical event P^*, capable of determining the occurrence of H_2, occurs or not.

Now, if this line of thought is correct, this means that H_1 is not the cause of H_2, because not every relevant scenario in which H_1 is present is a scenario in which H_2 is also present, and not every relevant scenario in which H_1 is absent is a scenario in which H_2 is absent. A given intervention on the putative cause will not determine the presence or absence of the effect; rather what will make a difference in this sense is whether the right lower-level, underlying physical states of affairs are in place. The upshot is that, while the interventionist picture outlined by Zhong does not force the epiphenomenalist conclusion upon the nonreductionist, it does not establish the non-epiphenomenalist conclusion either. This has two interesting corollaries. First, it should be noted that there is no awkwardness in maintaining that some higher-level event H_x does not have a cause, as all we have established so far is that, if H_x does have a cause, then that cause cannot be a lower-level physical one. But at the same time, because of *Determination*, this does not mean that the occurrence of H_x happens *by fiat*: H_x is (non-causally) brought about by whatever physical state of affairs it is determined by. Therefore, differently from physical states of affairs, it is not at all unusual for non-physical states of affairs to occur without a cause[36]. This, at last, is what makes the attribution of causal powers to higher-level

[36] Indeed, even though the interventionist account of causation does not conceive of causation as having inherently to do with energy, maintaining that non-physical states of affairs are caused by physical events may potentially lead to more awkward implications than maintaining that they are not, since non-physical states of affairs should not be conceived of in terms of matter-energy,

states of affairs redundant at best, and incorrect at worst. Sure, one can maintain that some H_x is the cause of some other H_y without this leading to any inconsistency. But the only way to motivate this claim would be via an appeal to our intuitions that, while rooted in long-established practices, would ultimately lack solid grounds, for those practices would be compelling only because of their being so widely endorsed, not because there is some robust reason for adopting them.

Let us now consider an illustrative scenario that will bring this whole discussion back to the 1PP condition on cognition. Suppose that some cognitive phenomenon *C* causes the occurrence of some other (perhaps non-cognitive) phenomenon *E*. If the 1PP condition on cognition is granted, then there will be some higher-level fact, *1PP*, associated with *C*. But *1PP* must be determined in its occurrence by some lower-level mechanism E_a. To keep things simple, we may grant that E_a does not supervene on anything else (it is a physical state of affairs analogous to P_1 above). Now, *E* may or may not possess properties that are determined by some lower-level state of affairs. In that case, since downward causation is ruled out by the interventionist argument, if we assume that there is a causal relation between *C* and *E*, *1PP* could not be the cause of *E*. Let us then grant, for the sake of the argument, that *E* involves some higher-level property or event *M*. By *Determination*, *M* must be determined by the occurrence of some lower-level property or event E_b. Since we are assuming that there is a causal relationship between *C* as a whole and *E* as a whole, E_b must have a cause. Not only that, but by *Closure* it must have a physical cause: E_a.

We now have a scenario analogous to the one I have been discussing above. E_a non-causally determines the occurrence of *1PP*, while it is the cause of E_b, which in turn non-causally determines *M*. Again, in the light of Zhong's interventionist argument it is not impossible to maintain that *1PP* causes E_a, but why should we do so? All we know is that *C* causes *E*, and this can be exhaustively accounted for by the causal interaction between E_a and E_b. What need is there to suppose that there also is a causal interaction between *1PP* and *M*? To be sure, we routinely do put forward claims of this kind, for example when we say that it is one's (first-personal-)desire to grab a cookie that causes one's grabbing the cookie. But, aside from this being what we routinely say, there is no more robust metaphysical motivation for taking this causal claim literally.

A possible response to my argument is that it is question-begging. If one starts by assuming that there is a causal link between E_a and E_b, and then proceeds to argue that since there is a causal link between these two variables there is no need to perform any intervention on *1PP* to see whether such an intervention would provide reasons to think that *1PP* is the cause of *M*, then one is effectively denying in advance that there can be a parallel, autonomous, causal chain going from *1PP* to *M*. But this is precisely what the interventionist strategy, at least in Zhong's case, purports to show. Thus, my argument unduly prevents the interventionist strategy from being played out: it focuses on one of the two components of it (E_a's causing E_b), and then discourages us from attempting to check whether the other component (*1PP*'s causing *M*) is in place too.

This reply would be, in a way, correct. It is true that my argument starts off with the causal relation between E_a and E_b. What the reply would neglect is that this is not the only nor the main reason why I maintain that we should not attempt to establish the parallel and autonomous causal relation between *1PP* and *M*. The reason why I maintain so is that another, non-causal sort of relation is in place, namely the determination relation between E_b and *M*. The presence of E_a causally determines the presence of E_b, which then non-causally determines the presence of E_a. Hence, if one were to argue that *1PP* causes *M*, it would follow that *M* is overdetermined[37]: on the one hand it is causally determined by *1PP*, on the other hand it is non-causally determined by E_b. While this does not entail any outright inconsistency, I maintain that parsimony considerations should lead us to prioritise one form of determination over the other. In fact, the causal determination between *1PP* and *M* cannot occur without the other form of determination, that between E_b and *M*, while the latter (non-causal) determination can occur if the other were not to hold. This goes against the spirit of interventionist causality: something is a cause of something else if it is a difference-maker, that is, if its being in one way or other is, *ceteris paribus*, uniquely responsible for something else's being in one way or another. The interventionist should not endorse this option.

[37] Notice, however, that this is not the sort of overdetermination that *Exclusion* bans, as it is not purely causal overdetermination.

4.3 Summary

Let us take stock. In the previous section, I have shown that the 1PP condition on cognition is a product of the anthropogenic approach to the elaboration of a mark of the cognitive. In addition to that, it is a nonreductionist condition. This means that the 1PP as a kind is not reducible (i.e. identical) to other kinds of things on which only a 3PP can be had, and this applies to all particular occurrences of phenomena which have a 1PP associated with them. But this nonreductionist reading makes the claim vulnerable to the famous exclusion problem, so that there is reason to think that the 1PP is really an epiphenomenon. This, in turn, is problematic if we keep (Dictum) fixed in the background, because that would mean that the 1PP's lack of causal powers would undercut the claim that it is necessary for cognition, as no epiphenomenon can be what characterises a real natural phenomenon (and cognition is assumed to really exist as such).

5. Abandoning Alexander's Dictum?

The bottom line of the previous discussion is that the 1PP condition on cognition is not a viable proposal. If (Dictum) is assumed, the exclusion problem makes the 1PP condition unsuitable as part of a mark of the cognitive, because the epiphenomenality of the 1PP would make it impossible for it to be a core component of what makes cognition the specific sort of natural phenomenon it is. However, there is one last attempt that one can make to rescue the 1PP condition. If (Dictum) were dropped, there would be no need to attribute causal powers to the higher-level properties or processes a MOC would be concerned with. In this section, I will argue that this is not an option, not only for a proponent of the 1PP condition, but also in general for a proponent of other nonreductionist anthropogenic proposals. The conclusion will ultimately be that, first, the 1PP condition cannot be saved, and second, that abandoning (Dictum) is not a promising avenue for saving nonreductionist anthropogenic marks of the cognitive in general either.

5.1 Alexander's Dictum and minimal naturalism

The main motivation for (Dictum) is that, provided that naturalism is on the right track, there are no mysterious forces governing the course of the events of the world.

Everything is ultimately grounded upon scientifically (physicalistically?) acceptable successions of states of affairs. Consequently, claiming that something is necessary for cognition (a natural phenomenon) seems to imply that such a feature is necessary because it plays some crucial causal role that cannot be eliminated if cognition is to occur, as every natural phenomenon needs to be fully accounted for in causal terms. In other words, there is a sense in which naturalism appears to justify the acceptance of (Dictum), and consequently of (1*)[38], so that cognition must be constituted by properties and mechanisms which are capable of causal interactions.

However, this conclusion may not be fully warranted, as it seems that an excessively restrictive understanding of naturalism is relied upon. Such a strong reading of naturalism may be unmotivated, as there are weaker alternative readings available. For instance, as proposed by Clavel Vázquez and Wheeler (2018, p.150), a more plausible reading of naturalism is what they call "minimal naturalism", which takes naturalism to just be the normative claim that, if there is a clash between some well-established scientific claim and some philosophical claim, there should generally be more pressure on philosophy to re-examine its product, rather than on science. Importantly, this does not mean that whenever an individual philosophical claim is at odds with any single scientific claim, philosophy has to give way. It only means that philosophers ought to be wary of neglecting the robustness of well-established scientific claims. Nonetheless, it may happen, at least in principle, that particularly strong philosophical considerations ultimately triumph over some scientific views.

In the light of this minimal naturalism, there would not be any obligation to endorse (Dictum). Surely, a great deal of what happens in our actual world is causally brought about, and it is possible to ascertain its existence in virtue of its causal powers. But this does not mean that certain properties cannot act as necessary conditions for natural phenomena while being incapable of exerting any causal power. In other words, one does not need to be committed to the view that natural phenomena like cognition cannot partially consist in the occurrence of epiphenomena. The worry one may legitimately have is that some conservative laws, in particular the undeniably well-established law of conservation of energy, would then be violated. But this need not be the case. For

[38] Recall that (1*) is the application of (Dictum) in the context of accounts of cognition: something can be claimed to be necessary for cognition only if it has causal powers.

presumably there would be a nomological connection between whatever higher-level property may be proposed as necessary for cognition and its lower-level physical realisers. Such connection would be such that, at least as a matter of nomological necessity limited to the actual world, whenever the material realisers of cognitive processes occur, certain epiphenomenal properties occur as well. Those causally effective realisers could not consequently occur in the absence of the relevant higher-level property. Interestingly, this counterfactual dependence would not rely on any causal influence exerted by the higher-level property as such. Rather, it would be a brute fact about our world, one that has nothing to do with the conservation of some physical quantity.

Let us see where we are at. I have pointed out that (Dictum) appears motivated by a strong reading of naturalism demanding, perhaps as a result of considerations about conservative laws, that everything in the world interacts with anything else in a causal, scientifically addressable way. However, more laid-back conceptions of naturalism, such as Clavel Vázquez's and Wheeler's minimal naturalism (or, in a way, Price's (2011) subject naturalism), only require that philosophical claims do not clash with comparatively more robust and empirically corroborated scientific claims. If one is to conceive of naturalism this way, the main point in support of (Dictum) could be dismissed, and with it the negative consequences that epiphenomenality would bring to a higher-level property claimed to be necessary for cognition. By showing that there exists a nomologically necessary connection between the realisers of cognition and the associated higher-level properties, one would be able to make the exclusion problem harmless for nonreductionist proposals of a mark of the cognitive.

5.2 Problems with the rejection of Alexander's Dictum

It seems, then, that the troubles for nonreductionist and anthropogenic proposals about the nature of cognition would be dispelled by the rejection of (Dictum,) and that, despite its plausibility, one may be inclined to do so if one endorses minimal naturalism. Upon closer inspection, however, this is not a move that a supporter of the 1PP condition can make. The endorsement of any metaphysical claim comes with costs, not just with

benefits. And, in the case of minimal naturalism, a supporter of the 1PP cannot pay the bill.

In the previous chapter I have argued that the 1PP, as I have characterised it, is remarkably similar to the (pre-reflexive) self-awareness much discussed in the modern phenomenological literature. This raises the question of whether the two really are different things, rather than one and the same thing. Of course, even if two notions are characterised in similar ways in different contexts, one should not immediately jump to the conclusion that there is not some respect under which these notions can be distinguished. But there is good reason to believe that, even though the 1PP and self-awareness turned out not to be the same thing, they are closely related enough to raise the suspicion that the 1PP may not have to do with cognition broadly construed as much as it has to do with consciousness instead. First of all, because self-awareness is meant to be one of the core phenomena associated with consciousness: it may not be sufficient for consciousness, but it seems close to being a necessary component of it. Second, the reason why having a 1PP associated with them is suggested to be a necessary requirement for certain phenomena to qualify as cognitive is that this is the result of a series of arguments vastly based on introspection. But such considerations based on introspection are made by conscious beings, and in particular, by conscious beings that are such not just in general, but also *while* introspecting. It would thus be highly surprising if consciousness were not to influence this process, at least to some extent.

It seems that either one can provide robust arguments in favour not only of the possibility, but also of the actual preferability of an understanding of the 1PP that makes it unrelated to consciousness, or one is forced to admit that the 1PP condition entails that cognition has to be conscious cognition. Since I do not see how the former option could be pursued, we are left with the second. But the second option, as I have argued in the previous chapter, immediately leads to the further claim that there can be no unconscious cognition. This ultimately makes the adoption of minimal naturalism impossible for the proponent of the 1PP condition. In fact, this claim would clash with the robust scientific consensus over the occurrence of unconscious cognitive phenomena. Hence, in the light of minimal naturalism, there should be significantly more pressure on the proponent of the 1PP condition to retract their claim. Not only that, but this condition cannot flaunt strong enough empirical evidence in its defence, as I have argued in the previous chapter.

As a consequence, the pressure on the 1PP condition is one that cannot be endured. The 1PP condition on condition must be dropped.

To recapitulate, minimal naturalism could save the 1PP condition on cognition from the exclusion problem, but the resulting picture would still be fatal for the condition. It is therefore time to finally dismiss it for good. However, the downfall of the 1PP proposal does not automatically concern all nonreductionist anthropogenic proposals. It may well be that, by embracing minimal naturalism, they can successfully escape the exclusion problem, thus remaining viable routes towards a mark of the cognitive. However, this happens not to be the case.

If (Dictum) stands or falls together with a form of naturalism stronger than minimal naturalism, then endorsing minimal naturalism would be tantamount to the rejection of (Dictum). Rejecting (Dictum) has the immediate benefit of disarming the epiphenomenalist conclusion of the exclusion problem. Doing so would in fact lift the ban on epiphenomenal properties as necessary for the occurrence of cognition. Since epiphenomena would be allowed in our ontology, there would be no immediate difficulty arising for someone who were to claim that some of those epiphenomena are necessary for cognition, even though cognition is a natural phenomenon, and even though cognitive processes are arguably not entirely epiphenomenal in nature. Nonetheless, one should ask whether it really is possible to feasibly claim that some epiphenomenon is necessary for a natural phenomenon like cognition.

Suppose that some epiphenomenalist condition (other than the 1PP condition) is claimed to be necessary for cognition. Whatever that condition presents as necessary for cognition, e.g. some property *Z*, cannot, trivially, cause the occurrence of anything. Can *Z* nevertheless be the effect of some other cause? Perhaps. But if it is, then it has to be such that any causal chain involving *Z* must end with it without exception, for *Z* could not be an intermediate link in a causal chain. From this, it plausibly follows that *Z* cannot be physical in nature, for any physical property can in principle be made to cause something, at least potentially. This is because if a physical object causally interacts with another physical object, the latter object will either exert an analogous causal influence on the former, or it will be at least able to causally influence other physical objects in a similar way. For instance, if some physical object is caused to increase its temperature, that object will be at least potentially able to cause other physical entities to increase their

temperature. Since this is not something that epiphenomena can do, and since one can imagine similar scenarios involving any physical property, it follows that epiphenomena in general, and *Z* in particular, cannot be physical in nature.

Now, cognitive phenomena are physically realised. Hence, in addition to *Z*, they must possess some physical properties. This means that cognitive phenomena considered in their entirety, rather than just with respect to *Z*, are not epiphenomena. What is emerging, then, is a picture where the manifestation of cognitive phenomena is causally insensitive to some property *Z* which is nonetheless maintained to be necessary for them. This raises the question: in which sense is *Z necessary* for cognition? The only way I think one can answer this question is in counterfactual terms. Our world is such that, were this "nomological dangler" (to use Smart's (1959) expression) not to occur, then, as a brute matter of fact, the cognitive phenomena associated with *Z* would not occur either.

While this picture is not an incoherent one, the credibility of a proposal for a mark of the cognitive that were to seriously endorse it would be dubious. First of all, I am not entirely convinced that no tension would arise, even against the background of minimal naturalism. To be sure, minimal naturalism is concerned with specific scientific claims, rather than with general scientific principles and methodologies. But this does not mean that it is in its spirit to allow for the proliferation of tensions with the scientific methodology, and I am inclined to think that such tensions would easily multiply if one were allowed to non-hypothetically posit something's existence despite its undetectability (as in the case of *Z*).

Second, the sense in which something is claimed to be necessary for cognition is not to be understood in exclusively counterfactual terms. Of course, that is possibly the biggest part of it, and it is a crucial interpretive key to make sense of claims of this sort. However, an explanation of why the nomological connection in question holds is also a non-negligible further component of such necessity claims. This component could hardly be supplied, were an epiphenomenon claimed to be necessary for cognition.

Finally, endorsing the above epiphenomenalist picture would make the quest for a mark of the cognitive utterly pointless. No understanding could be achieved from the discovery of a MOC, because the condition(s) constituting it would be presented as brute, inert facts, not as exploitable pieces of knowledge, nor as potential bearers of explanatory insights. There would be no use for an epiphenomenalist mark of the cognitive, for the

way cognitive processes unfold would be (causally) independent from the proposed conditions. Therefore, knowing that some phenomenon qualifies as cognitive in virtue of meeting the requirements laid out by an epiphenomenalist MOC would be nothing more than attaching a vacuous label.

To conclude, while there is nothing wrong with minimal naturalism, there is much that is wrong with the rejection of (Dictum). If the nonreductionist were to pit minimal naturalism against the (Dictum), that would result in an account of cognition that may not be incoherent, but which would lack credibility. In short, if the exclusion problem is not dispelled (and, currently, there is no conclusive argument to that effect), one cannot simultaneously go nonreductionist and attempt to elaborate an account of cognition. The question that remains to be answered is: are there any comparable difficulties associated with a reductionist approach? In the next section I will argue that yes, there are some difficulties, although not as troublesome as the ones I have been discussing so far. These difficulties appear to primarily concern proposals resulting from an anthropogenic approach, not from a biogenic one. Therefore, there is reason not to pursue an anthropogenic approach in researching for a MOC, and, conversely, there is reason to adopt a biogenic approach instead.

6. Reductionism and anthropogenic approaches

Let us start with the examination of what would happen if a mark of the cognitive were elaborated within a reductionist framework. Suppose that a necessary condition for the occurrence of cognition is identified, and this condition is expressed in terms of non-fundamental properties, that is, not in terms of the physical properties characterising the basic constituents of the world. Such non-fundamental properties must possess causal powers relevant to the occurrence of cognition. This is in the light of the requirement (1*) mentioned above, according to which something can be claimed to be necessary for cognition only if it possesses causal powers, and if such causal powers are somehow relevant to the occurrence of cognition. However, if this MOC is placed within a reductionist framework, the causal powers of the higher-level properties it mentions are exactly those of the lower-level properties they are reducible to. In other words, there would be no causal contribution that the higher-level properties could make in addition

to that made by the relevant reducing lower-level properties. This is because a plausible way to interpret reduction is as a sort of identity relation: if *x* reduces to *y*, then, numerically speaking, there is just one thing, which can be referred to as either *x* or *y*, so the causal powers of *x* are the same as those of *y*.

Before moving forward, it is worth mentioning a concern that one may legitimately raise, namely that, if reduction is to be understood in terms of identity, then reduction would be a symmetrical relation, because identity is symmetrical. But this is not the case. We want to say that the reduced entity does not, in turn, reduce its reducer. How can these facts be reconciled? This worry has been addressed by Van Riel (2013), who argued, in my opinion persuasively, that reduction generates intensional contexts, where the establishment of an extensional identity is not mirrored by an intensional one. I will not rehearse Van Riel's argument, but the upshot is this: two entities or states of affairs can be ontologically identical without being epistemically equivalent, so that it makes sense, from an epistemic point of view, to maintain that one reduces to the other but not vice versa. For instance, suppose that one were to successfully reduce a gust of wind to the movement of some air molecules. It is possible to sensibly maintain that what there is, ontologically speaking, is a single phenomenon, be it understood as a gust of wind or as the movement of air particles. But at the same time, it is also sensible to maintain, because of familiar theoretical considerations (explanatory power, breadth of scope, naturalistic posits…), that the movement of the air molecules, in virtue of its belonging to an intensionally different theoretical framework, takes precedence over the gust of wind as a higher-level entity, thus asymmetrically being the reducer of the gust of wind. The asymmetry of the reduction relation lies in the difference between the intensions associated with the theoretical terms in question, not in an extensional difference.

In short, reduction can be considered an epistemically asymmetric relation, while at the same time being thought to establish a numerical identity. In this sense, then, it is not problematic to retain the distinction between reducer and reduced entities, while simultaneously maintaining that such entities enjoy the same causal powers. Sameness of causal powers is guaranteed by numerical identity; asymmetry is guaranteed by intensional inequality. Now, I suspect that this scenario would turn out to be troublesome for a MOC proposing to take some higher-level properties or processes as necessary for cognition. This is because we normally lack a well-established account of the causal

powers making the case that those higher-level properties are involved in the occurrence of cognitive phenomena. Mentioning those higher-level properties or processes in a MOC would amount to nothing more than using a shorthand for referring to the lower-level processes doing the relevant causal heavy-lifting. Some may not have any qualms with this. After all, even if those causal powers are referred to by means of a shorthand, this does not mean that they are not also the causal powers of the higher-level entity. But I maintain that there is reason not to be content. In particular, in the absence of a satisfactory naturalistic account of the reduced higher-level properties in which their causal powers are specified, one would not be able to satisfy (1*), as the relevant properties' causal powers in play when cognition occurs would be unspecified. This scenario appears to most frequently occur in the context of anthropogenic, rather than biogenic, approaches, as it is not rare for them to put forward claims about the nature of cognition that involve phenomena or properties whose naturalised accounts are not available. In the rest of this section, I will elaborate on this point by taking as a case study Adams's and Aizawa's (2001, 2008) proposal, which not only qualifies as an anthropogenic MOC, but should also be interpreted as a reductionist one[39].

As I have been discussing in the first chapter, according to Adams and Aizawa (AA), cognition necessarily involves nonderived content, namely intentional content which a system as a whole, or some content-bearing part of it, possesses without this resulting from some stipulation, socially established practice, or, in general, influence exerted thanks to some previously existing intentionality. It seems uncontroversial to me that this claim is an anthropogenic one, not a biogenic one. Rather than being proposed as a result of the study of biological entities against the background assumption that cognition is just one among the various phenomena responsible for their survival, AA's proposal originates from the study of humans as paradigmatic cognisers. Hence, since this is enough to warrant considering AA's proposal as an anthropogenic one, I will not motivate this remark further.

The appearance of nonderived content in the debate over the MOC, and within the anthropogenic area specifically, is not entirely unexpected. Anthropogenic MOC's

[39] AA's proposal is here discussed as representative of reductionist, anthropogenic proposals, but the purpose of focusing on it in particular is purely illustrative. Should it be incorrect to interpret this specific proposal in reductionist terms, my general points about reductionist anthropogenic proposals in general would still hold.

generally establish a clear conceptual link between mind and cognition (Keijzer 2021), one which persisted through time because of the original derivation of the cognitive scientific project from the research on the mind, and since Brentano's times one of the most frequently endorsed marks of the mental is intentionality. It is thus not too surprising that nonderived content has been brought up also in the quest for a MOC. However, besides claiming that nonderived content is necessary for cognition, AA also suggest that, since accounts of natural phenomena ought, as the old saw goes, aim at carving nature at its joints, cognitive processes 'must be discriminated on the basis of underlying causal processes' (2001, p.52). Moreover, evidence of how AA take the lower-level details of cognitive processes to play a crucial role not only in discriminating the cognitive from the non-cognitive, but also in differentiating the various sorts of cognitive processes can be found later in the same page, where they argue that human chess-playing does not form a kind with computer chess-playing, the reason being not 'that the computer processes and the human processes are different; [rather, the reason is that], when examined in detail, the differences are so great that they can be seen not to form a cognitive kind'. These and similar remarks to be found in the fourth chapter of AA (2008)[40] are especially interesting for the present discussion, as they suggest, together with AA's endorsement of whatever attempt at naturalising intentionality will prove successful, that AA may welcome couching their claim on nonderived content in a reductionist conceptual framework.

This is a contentious point. One may object that it is preposterous to consider AA's view as a form of reductionism. After all, they also say that 'nothing in [their] mark of the cognitive says anything about the locus of cognition. […] Further, nothing about the kinds of processing in the brain conceptually, definitionally, analytically, or necessarily requires that they appear only with a brain' (2001, 53). This passage may be interpreted as speaking against taking them to be reductionists, perhaps because, given that they allow for nonderived content to occur in differently constituted systems, then AA cannot be taken to establish too strong a link between nonderived content and any particular underlying mechanisms such as reduction.

[40] 'The cognitive differs from the non-cognitive in virtue of the kinds of mechanisms that are involved' (2010, p.57), 'We think that developed sciences regularly categorize processes by reference to their underlying laws or mechanisms. Further, we think that cognitive psychologists aspire to do this as well and that they have discovered experimental methods that aid them in their work' (2010, p.60).

However, I believe that this objection would be misguided. AA do, in fact, establish such a strong relation. Indeed, they do not think that too coarse-grained criteria, such as behavioural or functional equivalence considerations, are satisfactory. What matters is that the right underlying causal processes are in place, abiding by what Wheeler (2010, 249) labels the "Adams-Aizawa distinctiveness principle": 'lower-level processes should be as distinctive as the higher-level processes they realize' (AA 2008, p.68). Hence, just like chess-playing as a higher-level phenomenon is not relevant for establishing whether a computer program is really playing chess in the same sense as humans do, it is not nonderived content *qua* higher-level notion, but nonderived content *qua* consisting of (numerically identical with) such and such information-processing mechanisms at a level lower than that at which content is ordinarily taken to belong to that proves necessary for cognition. This intensional inequality, paired with a numerical identity, warrants taking AA as having a reductionist understanding of nonderived content, even though they never explicitly commit themselves to reductionism.

To summarise, I think that AA's anthropogenic, higher-level proposal can be placed in a reductionist landscape. Or, at the very minimum, it should be "given the chance" to be placed in such landscape. In fact, it has to be understood in either reductionist or nonreductionist terms. But, given the exclusion problem, and the important negative consequences that it has on nonreductionist attempts at characterising cognition, interpreting AA's proposal in nonreductionist terms would be tantamount to saying that it should be dismissed as a viable proposal. Hence, for the sake of the argument, I will grant that it is not a hopeless proposal; but this will require pairing it with reductionism.

With that being said, I believe that, at least given the current state of our knowledge, the claim that nonderived content is necessary for cognition would be much weaker than one would have hoped, if paired with a reductionist background. It is true that for the purposes of their paper AA can rely on the broad favour their view enjoys ('There is […] a fairly broad consensus that cognition involves non-derived content' 2001, p.48; 'all we need presuppose is that cognition involves intrinsic content', p.49), without needing to endorse any specific naturalised account of intentionality. But, if one is to determine what makes nonderived content necessary for cognition on the basis of (Dictum), that is, if one is to determine to what extent cognition exists thanks to the occurrence of nonderived content, such an account is needed. This is because, if the causal powers of the higher-

level property of displaying nonderived intentionality are required to be identical to those of the lower-level states of affairs realising such property (as per reductionism), in the absence of a persuasive account of such causal powers it is unclear why exactly nonderived content would be necessary for cognition. In other words: why is nonderived content, with the causal powers shared with its corresponding underlying mechanisms, taken to be constitutive of cognition? Why would the causal powers of different mechanisms, such that they do not support the occurrence of nonderived content, not do the trick?

Of course, it is not AA's fault if we presently lack a solidly established account of what it takes to display nonderived content. Nonetheless, this issue is all the more pressing given the second claim AA advance, namely that cognitive processes should be individuated by their underlying lower-level causal processes. Without an established account of what these processes are, there is no way to individuate cognitive processes, and, as a result, it is hard to determine whether nonderived content really is a necessary condition for their occurrence.

To reiterate, even if the above considerations do not represent a conclusive argument against the adoption of a reductionist approach in the elaboration of an anthropogenic, higher-level MOC, I think that they at least make this possibility unpalatable, if only for the difficulties that it would bring upon one of the most prominent extant proposals. In particular, the plausibility often attributed to the claim that nonderived intentionality is necessary for cognition appears to stem primarily from higher-level considerations concerning the observation of actual or hypothetical (putative) cognitive systems. But, although one may still maintain that nonderived content is necessary for cognition while not being committed to reductionism, within a reductionist framework the support offered by these considerations is importantly weakened. The causal contributions relevant to the occurrence of cognition would be those of the reducers of nonderived content, not of nonderived content itself as a higher-level notion, for there is no evident way in which nonderived content involves causality. Of course, these contributions are the same, but in the absence of a satisfactory account of what the reducers in question would be, claiming that their causal powers are (among) what allows cognition to meet (Dictum) would be a rather uninformative claim. It would be nothing more than claiming that whatever accounts for intentionality in lower-level

terms is capable of playing a causal role, and, without specifying what such role may be, that is what constitutes cognition. This is far from providing a substantial constraint on what counts as cognitive, especially given that, even in higher-level terms, it is hard to go beyond a rough-and-ready characterisation of what nonderived content is.

It is worth stressing that AA's view has been discussed here not because I take issue with it in particular, but because it is representative of an anthropogenic proposal for a MOC couched in a reductionist framework. Anthropogenic approaches take human cognisers as their starting point, and they typically end up delivering some claim relative to the nature of cognition that involves higher-level properties. But such higher-level properties normally lack a naturalised account, especially one meant to reduce them to lower-level mechanisms. Hence, the sort of difficulties AA's proposal encounters generalise: if no successful reduction can be carried out, the causal powers of a higher-level property claimed to be necessary for cognition remain undetermined. Therefore, given (Dictum) and the related requirement for a MOC, (1*), it is unclear why the higher-level property in question would be necessary for cognition, in the sense of being constitutive of it, and hence contributing to the existence of cognition as a genuine natural phenomenon.

7. The rise of the biogenic approaches

Let us see where we are at. In section 5, I have argued that the exclusion problem plagues nonreductionist proposals for a MOC. This is because, if we accept (Dictum), one cannot advance an epiphenomenalist condition for cognition. Furthermore, undercutting the support for (Dictum) by adopting a minimal form of naturalism is not an available option for the 1PP condition, nor does it really help nonreductionist proposals in general (section 6). Hence, it seems that one should attempt to formulate a mark of the cognitive within a reductionist, rather than nonreductionist, framework. However, in section 7 I have highlighted how anthropogenic approaches struggle to offer a satisfactory proposal even in a reductionist context. This is because, on the one hand, one needs a naturalistic account of the properties claimed to be necessary for cognition in order to determine whether they are suitable candidates or not, i.e. whether they have causal powers. On the other hand, such naturalistic accounts are typically missing in the case of

the higher-level properties anthropogenic proposals are concerned with (e.g. the possession of nonderived content). As a consequence, it seems that anthropogenic approaches to cognition will not be viable until they will be able to advance solidly naturalistic conditions on cognition. To be clear, I do not intend to suggest that anthropogenic approaches to cognition should not be pursued *tout court*. Rather, I am pointing out that, currently, there is no satisfactory anthropogenic proposal, and that it is unlikely that there will be one in the foreseeable future.

It seems then that the only option available to someone intending to provide a MOC is to go down the biogenic path. Indeed, biogenic approaches to the issue appear to have a sizable advantage over anthropogenic approaches, as they typically offer conditions on cognition that are clearly naturalistic. Think of the various proposals mentioning allostatic control (Kiverstein and Sims (2021)[41]), sensorimotor coordination (van Duijn, Keijzer and Franken (2006)), or, more generally, adaptive behaviour (in a way, Keijzer (2021)). All these proposals are spelt out in terms of undoubtedly higher-level notions, but there is no deep difficulty in identifying the potential causal powers associated with them. For instance, were one to claim that adaptive behaviour is a necessary condition for the occurrence of cognition, one would be saying that whenever we observe physically instantiated observable behaviours whose outcome is the survival of the organism displaying them, that organism thereby qualifies as a candidate cognitive system. Causation, and in particular causal processes involving physical entities, is inherently built in such a characterisation, so I can see no reason to maintain that this is not a naturalistic proposal.

It seems then that the significant advantage that biogenic approaches enjoy over their anthropogenic rivals consists, primarily, in the fact that they typically do not need further philosophical work to be understood in naturalistic terms. This is unsurprising. After all, rather than taking their moves from considerations over human cognition, they take their moves from the observation of living organisms. Hence, the job of a biogenic project is proceeding in a bottom-up way from naturalistically depicted organisms lacking cognition to organisms that, instead, do manifest cognition. Naturalism is not something that needs to be brought into the picture later on, once some hypothesis about cognition is formulated; rather, it permeates the biogenic inquiry right from its outset.

[41] I will postpone the presentation of this particular proposal to the next chapter.

Of course, it is important to be clear about the extent to which a naturalistic conception of their objects of inquiry right from the start allows biogenic approaches to avoid issues related to the specification of the exact causal powers associated with what they end up claiming to characterise cognition. It may well be the case, in fact, that despite making uncontroversially naturalistic claims, some biogenic proposals mention complex higher-level phenomena whose exact realisers may not be fully understood yet, or which can be too difficult to exhaustively account for. Suppose again, for instance, that one were to suggest that some behaviour qualifies as cognitive just in case it is adaptive. The general idea is clear: if (and only if) the adoption of certain behavioural patterns is responsive to environmental stimuli in a way that affects the likelihood of the organism's survival, then the organism displaying such behavioural patterns is capable of cognition. However, what this exactly amounts to is unavoidably left to be specified contextually. Certain environment-sensitive behavioural patterns may be adaptive for some organisms and maladaptive for others; the very same kind of behavioural patterns may have a similar impact on the likelihood that two different sorts of organisms will persist over time, but for different reasons; or again, the mechanisms responsible for such behaviours may be unknown. It thus seems that the causal powers associated with the notion of "adaptive behaviour" vary across different contexts, and that they may also be (heavily) underspecified in many such contexts.

I do not think that this picture would speak against the biogenic approaches as a whole. This is for two reasons. First, in many cases the issue is practical in nature, rather than conceptual. Living organisms are complex entities, and it invariably is a hefty (and possibly Sisyphean) job to characterise in detail their mechanisms, despite the fact that frequently we have at least some general understanding of those mechanisms. Second, differently from the anthropogenic case, in the biogenic case the failure to specify the causal powers associated with the lower-level mechanisms responsible for (indeed, identifiable with, since we should be operating in a reductionist framework) the occurrence of the higher-level properties claimed to be necessary for cognition should not put the MOC in question on hold. This is because, for anthropogenic approaches, this amounted to not knowing what the causal powers associated with the relevant properties or phenomena were *at all*, while in the case of biogenic approaches this amounts to knowing only some of, but not all, such causal powers.

To see this more clearly, suppose that one adopts an interventionist account of causation. Recall that on this reading of causation, *x* qualifies as a cause of *y* if and only if there exists, at least in principle, some way to change *x* that leads to a change in *y*, while keeping everything else constant in the background. In the case of AA's suggestion that nonderived content is necessary for cognition, it is hard to see what such an intervention may be like. Of course, we may suppose that certain neural configurations or mechanisms are responsible for the instantiation of structures bearing nonderived content. But we do not have an account of what these neural configurations or mechanisms would look like. Therefore, it is not possible to make reasonable hypotheses about how we would have to tamper with the brain to examine whether nonderived content is the cause of certain putative effects. Some intervention on the brain may sort some observable effect, but we would not be any wiser about whether nonderived content was even one among the causes of such effect, especially because the *ceteris paribus* clause would be unspecified and unspecifiable.

On the other hand, in the case of our toy example of the claim that adaptive behaviour is necessary for cognition, we can easily imagine, at least sketchily, what an intervention on that would look like. Suppose that we were to tamper with a bat's echolocation. That would affect the bat's ability to hunt and, more broadly, to navigate the world, which in turn would affect its likelihood to survive. In this scenario, no mention of the exact mechanisms enabling the bat's echolocation is made, and yet we could reasonably maintain that echolocation has causal powers: in its absence, its putative effects (some of the bat's behaviours) would be altered. More specifically, we would be able to determine that echolocation is at least partly causally responsible for the bat's ability to survive. We could thus attribute *some* causal powers to one of the ways in which bats engage in adaptive behaviour. Therefore, given that at least in some cases (indeed, in most cases) it is possible to test and ascertain the causal powers of some coarsely construed parts of the realisers of a particular instance of adaptive behaviour, there is no conceptual difficulty, in this respect, in maintaining that adaptive behaviour is important for cognition.

To summarise, biogenic approaches have a crucial advantage over anthropogenic approaches, that is, they typically do not struggle with the specification of at least some causal powers associated with what they claim to be necessary for cognition. Once again,

it is important to stress that this is not an immutable state of affairs. It is possible that one day there will be an agreed upon naturalistic way to understand the sort of phenomena that anthropogenic proposals are concerned with (e.g. reasons, nonderived content, perhaps even the 1PP). But until then, we should go down the biogenic path. In the next, final chapter I will examine a currently popular view that stems from a biogenic approach, to illustrate what going down the biogenic path would look like.

Chapter 6[42]

Introduction

I have concluded the previous chapter by arguing that biogenic approaches have a substantial advantage over anthropogenic approaches, and for this reason they should be pursued by those interested in understanding what cognition is. More specifically, beginning one's inquiry on the MOC from the observation of biological systems typically results in the individuation of some characteristics of cognition that can be unproblematically understood in naturalistic terms. Thus, the adoption of reductionism, combined with the assumption of Alexander's Dictum (according to which some property can be claimed to be necessary for cognition only if that property possesses causal powers), would not lead to any especially troublesome consequences for biogenic approaches. This is because, even if we are not able to spell out in detail *all* the causal powers that such properties are supposed to possess, we can nonetheless at least indicate *some* of them. And this is enough to legitimise claims suggesting that those properties are necessary for cognition, although, of course, this has no bearing over the actual truth of those claims.

On the other hand, anthropogenic approaches typically (but not invariably) result in the suggestion that cognition necessarily requires certain properties or states of affairs that at present lack a clear naturalistic account. As a result, within the same reductionist framework, Alexander's Dictum would cast doubts over the viability of anthropogenic proposals of this kind, for it would not be clear what causal powers the relevant properties would possess.

In the light of these considerations, I have concluded the last chapter by recommending the adoption of a biogenic approach in one's search for a MOC. The issue is now to more precisely pin down what a biogenic approach would look like. This is the task that I will undertake in this final chapter. In particular, I will discuss what arguably is the most widely approved biogenic conceptual framework in the literature, namely the

[42] A version of this chapter has been recently published in its entirety in *Synthese* (Pisano, 2023).

Free-Energy Principle (FEP)[43], which can be understood as a particular form that the more general predictive processing framework has assumed over the past fifteen years or so, and which establishes an important continuity between life and cognition. It is however important to point out that, properly speaking, the FEP is not a theory of cognition. Rather, it is the broad theoretical pool, concerned with life in general, from which some theories more specifically targeting cognition can emerge (e.g. Kiverstein and Sims (2021)).

Before turning to the FEP proper, a brief note on its relation to the predictive processing research programme. Predictive processing is one of the most popular axes along which research on cognition is developed (Hohwy (2013); Clark (2016)). The core insight driving this research programme is that cognitive systems operate in probabilistic terms, formulating predictions about their environment and then adjusting them based on whether their expectations are met or not. Among the various elaborations of this conceptual pillar is the FEP (see Friston, Kilner and Harrison (2006), Friston (2009, 2010, 2012, 2013)), which, in a nutshell, states that adaptive systems strive to keep their free-energy (a proxy for surprise, which is an information-theoretical notion) at a minimum, by making the case that they remain within a certain range of (unsurprising) states enabling their survival. In particular, this is done by behaving in a way that approaches optimal Bayesian inference. Based on a prior probabilistic distribution linking environmental states of affairs to the sensory states the system may enter in because of them, as well as on the basis of the actual sensory states the system enters in as a result of environmental influences, adaptive systems can try to act on their environment so as minimise the likelihood that they will enter in unsuitable states for their own survival.

It is then with the Free-Energy Principle that I will be concerned in this chapter. In particular, I will examine the question of whether we should adopt a realist or an instrumentalist approach to the models that are crucially involved in the study of adaptive systems. In doing so, I will make use of some insights coming from the literature on scientific modelling to show that we should indeed embrace instrumentalism. However, this will have interesting consequences for attempts at exactly characterising the nature of cognition taking the FEP as a starting point.

[43] Whether the FEP is, properly speaking, a principle or a conceptual/mathematical framework is not too important here, although it is indeed an interesting issue. Here, I will refer to it as a principle that is meant to apply to a conceptual-mathematical framework.

I will proceed as follows. I will begin (section 1) by offering a general characterisation of the FEP. The purpose is to present as informally as possible the main ideas associated with it, as well as the theoretical tools it employs. Then (section 2), I will argue that the models involved in FEP-theorising should be plausibly understood as being isomorphic to their targets (although I will remain non-committal with respect to the structuralist view of scientific representation in general, especially when based on isomorphism). This will allow me (section 3) to turn the criticisms moved against isomorphism-based accounts of representation towards the modelling practice involved in the FEP. That is, maintaining that FEP-models represent their targets as they are, in a realist sense, is unwarranted. This is because the failure to establish an isomorphism between a model and its target leads to a failure on part of the former to represent the latter, and because it is highly unlikely that FEP-models are ever isomorphic to their targets due to unavoidable design choices (driven by explanatory interests) involved in modelling practice. Consequently, while FEP-models can be empirically adequate, we should refrain from interpreting them in a realist way and go instrumentalist instead.

Finally (section 4), I will return to the issue of finding a MOC, and I will consider what implications my argument in favour of an instrumentalist reading of the FEP-models has for attempts at making use of the FEP to elaborate an account of what cognition is. My conclusion is that we should not dismiss accounts of cognition based on the FEP, as they may still be informative and further our understanding of the nature of cognition. Nonetheless, the prospects of settling some of the philosophical debates that sparked the interest in having a "mark of the cognitive" are not good.

1. The Free-Energy Principle

I will get things started by presenting the framework I will be concerned with in this chapter. This will be an informal introduction, whose purpose is to outline the core concepts and ideas constituting the Free-Energy Principle (FEP), with a special focus on the way models are made use of in this framework. While the details of the FEP are quite technical, my presentation will be completely informal, thus unavoidably imprecise at times. But before starting, a couple of notes. The labels "active inference" and "Free-Energy Principle" are interchangeable, and are equally frequently employed in the

literature. Here, I will tend to use the label "active inference" to refer to the sort of processes implementing the FEP, while I will use the label "Free-Energy Principle" for the conceptual framework as such.

Furthermore, the exact understanding of the epistemic status of the FEP is an object of debate (see, for instance, Andrews (2021) and Hohwy (2021)). The consensus is that the FEP adopts a "principle first" approach (Van Es and Kirchhoff (2021, p.6623)) in offering an array of mathematical instruments to conceptualise and describe self-organising systems and their behaviours. However, the FEP does not appear to be a proper scientific theory; rather, it seems to be better conceived of as a principle, or as a mathematical framework. In what follows, I will refer to the FEP as a theory exclusively when I will be discussing the FEP *qua* coupled with some (contextually unspecified) process theory concerning its implementation. Otherwise, I will refer to the FEP more neutrally as a "conceptual framework".

Finally, one important caveat is in order. As mentioned in the introduction, the FEP is tightly related to predictive processing, and, in a way, it can be thought of as a particular development of it. Accordingly, just as there exist numerous versions of predictive processing (predictive coding, prediction error minimisation…) which differ from one another in a number of respects, various readings of the FEP are available. Some are neurocentric, some are not; some are representationalist, some are not. For instance, Hohwy (2015) endorses a neurocentric, representationalist reading of the FEP, while Kirchhoff's and Kiverstein's (2019) is a non-neurocentric, non-representationalist reading. Here, my presentation of the FEP will be largely based on the latter approach, which is enactivism-flavoured[44]. With that being said, for the purposes of the argument I will make, it does not really make a difference what process theories are coupled with the FEP, nor whether one has (non-)representationalist or (non-)neurocentric inclinations;

[44] It is important to note, however, that I am not committed to the *actual* compatibility of the FEP and enactivism. There is an ongoing debate in the literature over the possibility to combine the FEP and enactivism (Allen and Friston (2018); van Es and Kirchhoff (2021); Raja et al. (2021); Di Paolo et al. (2022)). This, however, will not concern us for present purposes, as the characterisation of the FEP presented in this chapter will not make use of any technical notion coming from the enactive literature, nor from that on autopoiesis. Thus, its being enactivism-flavoured ought to be understood as pointing at features such as embodiedness, non-neurocentricity and non-representationalism, which are taken to be part of what motivates attempts at reading the FEP in properly enactivist terms, but which are not sufficient for establishing any robust relationship between the FEP and enactivism or autopoiesis.

whenever the need to be explicit about such commitments should arise, this will be made clear.

1.1 The Free-Energy Principle: an overview

The FEP can be seen as a specific development of the currently popular view that cognitive systems are predictive systems. That is to say, cognitive systems can be understood as approximating optimal Bayesian machines. As such, they formulate hypotheses in accordance with probability theory about their environment, which also includes the cogniser's internal states that are not involved in cognition. These hypotheses are what inform the cognising organism's perceptions and, consequently, actions.

As it happens, there are various ways to interpret the Bayesian inferences cognitive systems perform, depending on how literally one takes them to occur (on this point, see Kirchhoff, Kiverstein and Robertson (2022)). One may adopt a fully literal reading, and claim that cognitive systems engage in explicit, personal-level inferences. However, this reading is likely to lend itself to all sorts of criticisms, among which some are analogous to the well-known homunculus fallacy. As far as I am aware, this reading is not endorsed by many, if any, scholars working in the field, and I will accordingly leave it aside.

Alternatively, one may take a weaker, realist stance, and maintain that cognitive systems do engage in Bayesian inference, but not in a personal-level or explicit sense. In Hohwy's (2015, p.17) words: 'The brain itself does not, of course, know the complex differential equations that implement variational Bayes', but nonetheless 'the brain is literally Bayesian in much the same sense as the heart is literally a pump' (*ibid*.). This is the reading I have in mind in the current presentation of the FEP, and to which the instrumentalist stance I will argue for will be recommended as an alternative.

What the FEP adds to the predictive processing picture is a general principle that guides the Bayesian inferences corresponding to the various hypotheses a brain (or an organism, if one opts for an embodied reading) formulates: activities based on predictive processing tend to minimise (variational) free-energy[45]. Free-energy is an information-

[45] It can be observed that this formulation is ambiguous between two readings. According to the first, variational free-energy minimization is construed as the objective function that guides the drawing of approximate Bayesian inferences. According to the second reading, free-energy minimization is a sort of imperative that living and cognitive systems need to abide to in order to persist. I believe that both readings are viable, but the latter is more appropriate in this context.

theoretical quantity which poses an upper bound on surprise[46]. That is, given the *actual* states a cognitive system enters in because of the environmental data, and given the *predictions* concerning the states a cognitive system would enter in because of the expected environmental data, the measure of the mismatch between the actual and the predicted internal states, i.e. the free-energy, is always greater than surprise. Surprise, in turn, is a quantity closely related to Shannon entropy, as 'on average, entropy is the long-term average of surprise' (Parr, Pezzulo, and Friston (2022, p.48)). Therefore, the FEP maintains that cognitive activities purport to minimise surprise, hence entropy, not directly, but by minimising its maximum value, which is set by free-energy.

The motivation behind the FEP is, in brief, the following. Shannon entropy (an information-theoretical quantity) is formally similar to thermodynamic entropy (Colombo and Wright (2021, p.S3472)). Thermodynamic entropy, in turn, can be generally understood as a measure of disorder, and it naturally tends to increase, as per the second law of thermodynamics[47]. But living organisms, in order to remain alive, need to be organised in certain specific ways depending on the sort of organisms they are. Therefore, organisms need to "resist" this tendency towards disorder. Based on the FEP, engaging in cognitive activities is one way to do so.

Now, how do cognitive systems minimise their free-energy? They engage in active inference, which consists of two complementary sorts of processes that do not need to take place sequentially; on the contrary, they can and typically do occur in parallel. On the one hand, systems update the probabilities upon which their predictions are based. In other words, while the prior probability distribution of the supposed causes of their observations is represented by the generative model, if confronted with surprising sensory inputs, cognitive systems modify their recognition model, which represents the observationally informed posterior probabilities of the causes of their observations (see Ramstead, Kirchhoff and Friston (2020)). On the other hand, cognitive systems also minimise free-energy by actively modifying their environment through action, consequently changing the inputs received so as to more closely approach the expected

[46] The surprise associated to some observation, or, more precisely, to the sensory states a system enters in as a result of being influenced by its environment, is the negative log probability of that observation.
[47] It is rightly customary in presentations of the FEP to point out here that what is properly involved here is not the second law of thermodynamics, but the fluctuation theorem.

ones[48]. In this sense, cognitive systems are engaged in self-fulfilling predictive processing which consists in selectively sampling their environment (see Hohwy (2016)). A circular dynamics is then in place. Cognitive systems formulate predictions about their environment based on their previous information, and if their expectations are not matched by the incoming data, they modify themselves and the environment from which the surprising data come, so that their subsequent predictions are less likely to clash with later data.

Some crucial remarks are in order. First, the reason why self-organising, adaptive systems (and, consequently, cognitive systems) are construed under the FEP as attempting to minimise free-energy, rather than surprise directly, is that those systems cannot assess how surprising the states they enter in as a result of environmental inputs are. They just are in those states. A system does not represent its predictions to confront them with the environmental data in order to measure its corresponding surprise, as that would be a computationally intractable task. Rather, systems *embody* their recognition model: their internal states are interpreted within the FEP conceptual framework as being the predictions themselves, rather than representing them. In this sense, organisms are engaged in a process of self-evidencing. This means that, being themselves predictive models of their own environment, by gathering confirming evidence in favour of their predictions they correspondingly gather evidence in favour of themselves being good models of their environment (in accordance with Conant's and Ashby's (1970) good regulator theorem)[49].

Relatedly, the way systems update their recognition model (i.e. the way they change so as to embody different expectations) and act upon their external environment to modify it does not follow any higher-order rules. Their generative model, i.e. the patterns followed in reacting to surprise and consequently engaging in active behaviour is not the sort of thing that cognitive systems consult to obtain guidance over their behaviours. The

[48] It is worth mentioning that while up to this point I have been talking of free-energy having *variational* free-energy in mind, i.e. "actual" free-energy, when it comes to the active part of active inference the relevant sort of free-energy is *expected* free-energy, that is, the free-energy that is expected to be associated to the future states the system will enter in as a result of different behavioural policies. With that being said, I will keep using the generic term "free-energy" in the rest of this chapter.

[49] It is worth emphasizing that this is a decidedly embodied, non-representationalist way of putting this point. Non-embodied, representational alternatives are available in the literature (e.g. Hohwy (2015, 2016), Gładziejewski (2016)). The main difference is that, according to them, organisms (and cognitive systems) do not embody, or are not themselves, their own models; rather, they *have*, or *make use of* those models.

generative model can only be inferentially abstracted away from the actual behaviours adopted by cognitive systems without it being at the cognitive systems' immediate disposal.

In short, cognitive systems are interpreted, within the theoretical framework of the FEP, as embodying a recognition model, and they are said to entail a generative model (Ramstead, Kirchhoff and Friston (2020)). The minimisation of surprise, the ultimate intended effect of cognitive activities, is not what cognitive activities tend to do *per se*, because what determines surprise is not available to cognitive systems. However, what free-energy depends upon, namely the recognition model and the sensory states a system enters in as a result of certain environmental data, *is* accessible to cognitive systems, which can then try to minimise it. And, since free-energy represents an upper bound on surprise, i.e. it constrains the maximum value of surprise, minimising free-energy has the consequence of indirectly minimising surprise.

1.2 Markov blankets and generative models: models in active inference

So far, I took the distinction between the states of a cognitive system and the environment (the external states) for granted. However, it should be clear that this separation is too important for the FEP to leave it unaddressed. This is because quantities like free-energy, surprise, and the probabilities involved in the generative model all need to be quantified on the basis of the internal states of the cognitive system and of what is part of the environmentally sourced data. The separation between internal and external states is mathematically handled with the help of Markov blankets, which are derived from Pearl's (1988) notion of a Markov boundary (a Markov blanket which does not contain other Markov blankets as its subsets).

The notion of a Markov blanket is a graph-theoretic one, and it was 'introduced as a way of separating a node in a Bayesian network from other nodes in the network' (Menary and Gillett (2022, p.41)). Hence, *per se*, Markov blankets are a purely formal tool used in the study of artificial Bayesian networks, and they do not straightforwardly correspond to any real-world state of affairs. This has led many[50] to forcefully contest the

[50] Bruineberg et al. (2021); Menary and Gillett (2022); but also Facchin (2021), although in this case limited to the extent in which Markov blankets can be used to settle disputes over vehicle externalism.

use that is made of Markov blankets in the literature on the FEP. For, a clear move is made from the original, epistemic use of the Markov blanket formalism in a non-empirical context, in the direction of a metaphysically committed use in an empirical context. In other words, what is contested is the legitimacy of taking this formal device to apply to the real world, in the sense of being able to fully account for the demarcation of the boundaries of self-organising systems purely in virtue of formal characteristics. If Markov blankets are to be used in this way, some interpretive assumptions are required, and such assumptions cannot be extrapolated from the Markov blanket formalism as appearing in its proper graph-theoretic domain.

I am sympathetic to these criticisms. However, for present purposes, I will assume that it is conceptually legitimate to make use of the Markov blankets formalism in the way the literature on the FEP does. To be clear, this is not to say that I am assuming a realist interpretation of the Markov blankets. Rather, I am conceding for the sake of the argument that it is not a category mistake to maintain that Markov blankets are what delimits the boundaries of self-organising systems.

The core idea underlying the Markov blanket formalism, as employed in the context of the FEP, is that something is part of a living system in so far as it plays a statistical role in shaping the later developments of the system. Consider a set of objects, {1, 2, 3, 4, 5}[51]. Let us suppose that some of its elements are conditionally dependent upon some other elements. That is, depending on the states that the latter are in, the former have a varying chance to enter in certain other states at a later time. In particular, suppose that: 3 is probabilistically relevant for 4 (so that depending on the state 3 is at a certain time, 4 will have a certain probability of entering in some state rather than another at a later time); 1 is also relevant in an analogous way for 4; 2 is relevant for 3; and 4 is relevant for 5. We have the following situation:

[51] This presentation of the core idea behind the Markov blanket formalism is largely based on Clark's (2017) and Hohwy's (2017) illustrations.

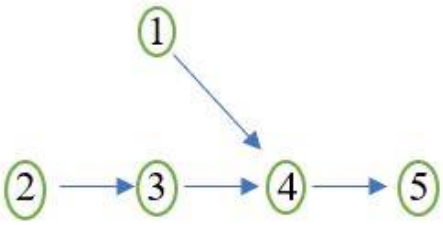


Fig.1

Let us now define 'the parents of a variable X [as] the variables whose directed connections lead immediately to X; [and] the children of a variable X [as] the other variables to which the X leads immediately through its directed connections' (Facchin, (2021, p.5)). In our scenario, the parenthood relationships are as follows: 2 is a parent of 3; 4 is a parent of 5; 1 and 3 are parents of 4.

With the needed terminology in place, it is possible to define a Markov blanket for each one of these elements as a set of nodes that makes a given node of the model in question conditionally independent from all the other nodes in the model. This set includes the parent(s), the children, and the parents of the children of the target element, and it is such that all the other elements of the model are probabilistically uninformative with respect to the task of determining the states of the element in question. For example, the Markov blanket associated with 4 in the illustration above would be constituted by its parent nodes 1 and 3, and by its child 5, while the blanket associated with 3 would comprise its parent 2, its child 4, and its child's parent 1.

Now, how is all this employed within the FEP framework? As anticipated, this formalism is used to separate the internal states of a cognitive system from its external, environmental states. Accordingly, the node upon which a given Markov blanket is centred represents the internal states of the system itself, while the nodes constituting the relevant blanket constitute the blanket states of the system. These states are still part of the system in question, but they are not statistically separated from the environment. In particular, two sorts of blanket states can be individuated: the sensory states, which are statistically dependent on the environment and on which the internal states of the system are, in turn, dependent; and the active states, namely those which statistically depend on the internal states of the system, and upon which the environment is statistically dependent. To illustrate this with one example from above: if the internal states of the

system are represented by node 4, the sensory states would be represented by nodes 1 and 3, while node 5 would represent the active states of the system.

So, a cognitive system is represented by means of the Markov blanket formalism as constituted by its internal states and by its blanket (sensory and active) states. It is at this point that we can see how the predictive processes involved in the FEP enter the scene. The bulk of a cognitive system is not directly statistically related to its environment, as it is not directly acted upon by the environmental states of affairs, nor does it directly act upon them. Furthermore, given that the boundaries of cognitive systems are determined by a Markov blanket, whatever lies outside those boundaries is not immediately available to the cognitive system as a whole, as it is something "other", external. For this reason, cognitive systems can only make educated guesses about what actual states of affairs cause them to enter certain states (about the inaccessible generative *process*), as well as about what sort of impact on their environment certain courses of action (*policies*) will have. Such predictions are based on: prior probability distributions concerning the likelihood of various external states of affairs; the statistical correlations between those states of affairs and the resulting sensory states caused by them; the expected sensory states which will be later on caused by the environment depending on the policies undertaken. Having all this in mind, one can say that the actual behaviours adopted by cognitive systems *entail* a generative model, that is, a model which represents the aforementioned factors that ultimately lead to the observable actual behaviours.

Consider the following example. I am sitting next to a pond, on a summer evening. At some point, I slap my arm, then reach for some mosquito repellent. Based on my observable behaviour (my active states), in the light of the FEP, one can infer that I have entered some surprising sensory state such as one caused by an insect bite, and I have undertaken action so as not to enter in a similar undesirable state later on. In this sense, I, as a cognitive system, with my observable behaviours, can be said to entail a generative model. What would the generative model include in this case? Well, first of all a probability distribution concerning my sensory states. Some of them are not harmful, hence they are not associated with much increment in entropy, and, therefore, carry little surprise (recall that entropy is the long-term average of surprise). Some others, such as a stinging feeling, are instead surprising. Second, a likelihood distribution concerning the

potential causes of my sensory states; in our example, it is unlikely that a tiny dart shot with a blowpipe hit me, while it is much more likely that a mosquito bit me.

Hence, from the previous scenario one can infer a model of the mechanisms that unfold while I engage in my observed behaviour. Certain sensory states are surprising, they are more likely to be caused by mosquito bites than by tiny darts, and I believe that by using mosquito repellent I will not enter such states anymore in future. That is, applying mosquito repellent is a good (expected) free-energy minimizing strategy that will allow me to reduce future surprise. This is the sort of information that a generative model[52] contains: a specification of how the cognitive system takes the world to be, of what counts as a surprising state, and of the action policies apt for future surprise minimisation.

A question that has been in recent years discussed (e.g. Colombo, Elkin and Hartmann (2021); Van Es, (2021)) concerns the status of the models involved in active inference. Are generative models just instrumentally conceivable theoretical constructs which FEP-scientists make use of, which do not correspond to anything cognitive systems really avail themselves of? Or are they instead to be interpreted in a realist way, so that, as Parr, Pezzulo, and Friston (2022, p.172) suggest, the scientist's task is that of 'recover[ing] the parameters of the generative model that a subject's brain uses to produce behavior – the *subjective* model […by using their…] own generative model (of how the subjective model produces behaviour) – the *objective* model'? I will address this question in section 3. But before doing so, I will need to explain what notion of model seems to be at work within the FEP.

2. Models and isomorphisms

In the previous section I have presented the central ideas of the FEP, as well as the sorts of models of self-organising systems that are constructed based on the theory. Namely, living (and cognitive) systems are said to entail a generative model, i.e. the system's statistical representation of how the sensory inputs are generated as a result of

[52] Notice that I am here talking only about generative models rather than recognition models. The difference between the two is, again, that the former have to do with prior probabilities, while the latter with posterior probabilities. For present purposes, focusing on generative models and leaving recognition models aside will be of no consequence.

the interaction with the external environment. Such systems are statistically separated from their environment via a Markov blanket, comprising the sensory states the system enters in because of the environment's influence, as well as the active states the system enters in to manipulate the environment itself. Here, I will lay the grounds for the subsequent examination of the way the FEP-informed models are to be understood. Specifically, I will introduce the structuralist view of scientific theories, dating back to Van Fraassen (1980) (and notably discussed in Van Fraassen (2008)) and I will focus on the most relevant aspect of the theory for my purposes, namely the idea that scientific models relate to their targets via some morphism.

2.1 The structuralist view

One of the broadest topics from the general philosophy of science is the issue of scientific representation: how does it work? Is it different from other forms of representation, such as artistic representation (see Callender and Cohen (2006))? Of particular interest for our purposes is the structuralist approach, which is also sometimes referred to as the "mapping account" (Pincock (2004); Nguyen and Frigg (2021)). The core idea behind this conception of scientific representation is that scientific models represent their target phenomena by being similar to them in a specific way, namely by being related to them via some morphism. That is to say, insofar as models can be conceived of as structures (set-theoretical entities composed by a domain of elements and a set of relations defined over that domain), they can represent their targets if some morphism, i.e. some structure-preserving mapping, exists between the model and the target.

Two important remarks are in order. First, morphisms are functions, and as such they can only be defined over mathematical objects. Properly speaking, then, models are not morphic to their target phenomena *qua* natural entities, but to the mathematical structures encoding the data concerning such phenomena (see Nguyen (2016)). Second, depending on their exact properties, morphisms can be of different kinds: there can be isomorphisms, partial isomorphisms, homomorphisms, partial homomorphisms... For reasons that will be made explicit in a moment, I will here focus only on isomorphisms.

Let A = < D; P^n_j > and B = < E; Q^n_j > be two structures, respectively, a model and its target system, where D and E represent the domains of, respectively, A and B, and P^n_j and Q^n_j represent the n-places relations on, respectively, D and E. A function $f: D \rightarrow E$ is an isomorphism just in case two conditions are met. First, f is a bijection, so that no two elements of D are mapped on the same element of E, and for each element of E there is an element in D which is mapped on it. Second, f is relation-preserving, so that, for any n-tuple $(x_1, \ldots, x_n) \in D$, $P^n_j [x_1, \ldots, x_n]$ iff $Q^n_j [f(x_1), \ldots f(x_n)]$, and for any n-tuple $(y_1, \ldots, y_n) \in E$, $Q^n_j [y_1, \ldots y_n]$ iff $P^n_j [f^{-1}(y_1), \ldots f^{-1}(y_n)]$. Informally, this means that the model A univocally represents all and only the elements of the target system B, and all and only the relations existing among elements of the model are associated with relations among the elements of the target system. In other words, model A is an accurate and complete representation of the target system B.

Now, I maintain that there is reason to take the structuralist view in its isomorphism-based guise to be the sort of account that best describes the modelling practice FEP-theorists engage in. When constructing an *objective* generative model for some cognitive system, FEP-theorists gather data about their target system's behaviour. Then, they construct a mathematical model which purports to describe the *subjective* generative model[53] used by the cognitive system in engaging in the active inference processes leading to the observed behaviour. Such objective generative model can be understood as a structure whose domain's elements stand for the parts of the system whose states are taken to be relevant to the processes in question. Moreover, the relations defined over such domain can be understood as corresponding to the statistical correlations the system takes to exist among the elements of the domain. Finally, given that both the objective and the subjective generative models are mathematical structures, there is no threat of a category mistake arising from attempting to establish an isomorphism between them.

Before moving on, a couple of points need to be addressed. The first is a brief caveat concerning the scope of my claim. I am not claiming that the structuralist account of scientific representation is, in general, correct. That is to say, nothing in my argument relies on structuralism correctly identifying the necessary and/or sufficient conditions for the occurrence of scientific representation. The claim I am instead making is the

[53] Notice that this is not the same as the generative *process*, as the generative process is what the subjective generative model purports to reconstruct, but it is not what the objective generative model is constructed to capture.

substantially weaker one that, even if structuralism is not the ultimately correct account of scientific representation, it does seem to reflect at least the modelling practice involved in the FEP.

The second point is lengthier. According to the structuralist view, a model represents its target only if there exists a morphism among the structures corresponding to the two of them. But what I have sketched above is a way to interpret objective and subjective generative models as structures, without saying anything specific about the morphism allegedly mapping one onto the other. Furthermore, I have anticipated that I would have taken isomorphism to be especially relevant for our purposes, as opposed to other, weaker, morphisms. Why is that? The answer can be extrapolated from works such as Hohwy's (2016) and Kirchhoff's and Kiverstein's (2019, 2021) which discuss the issue of how to draw the boundaries of cognitive systems via the Markov blankets formalism.

2.2 Isomorphism and FEP-models

Isomorphism is a strong kind of morphism, as it requires that there is a relation-preserving bijection among two structures. Because of this, it has been noted that isomorphism-based theories of representation struggle to account for misrepresentation (Suárez (2003)). By misrepresentation, I mean a representation which is inaccurate (a representation possessing features which are not possessed by the target) or incomplete (it fails to represent some features of the target), but which nonetheless counts as a representation of its target. This inability to accommodate misrepresentations is problematic if isomorphism is meant to be the mapping upon which scientific representational processes are founded. In fact, it is not unusual that scientific models are incorrect in one way or another without losing their representational power as a result, as they instead should, were their powers based on isomorphism.

To respond to this worry, some have attempted to frame structuralist accounts of scientific representation not in terms of isomorphism, but in terms of weaker morphisms, such as partial isomorphism (e.g., Bueno (1997), Bueno and French (2011)) or homomorphisms (Bartels (2006); for a critical discussion, see Pero and Suárez (2016)). The idea guiding these alternatives to isomorphism is to enable structuralist accounts of representation to successfully deal with misrepresentation, that is, to allow models to

maintain their representational characteristics despite inaccurately or incompletely depicting their targets. Adopting a morphism weaker than isomorphism may afford more flexibility in accommodating misrepresentation, which is beneficial for accounts of scientific representation. However, there is evidence coming from discussions over FEP-models suggesting that isomorphism proper is what is relevant in that context. I have specifically in mind debates over the question of whether for each cognitive system there exists a unique Markov blanket enclosing it.

Recall that the task of a FEP-researcher, according to some of the most prominent upholders of the view, is to construct an objective generative model, which is meant to reflect the subjective generative model giving rise to the observable free-energy minimising behaviours a cognitive system displays. This suggests, as Clark (2017, p.12) pointed out[54] and as also Parr, Pezzulo and Friston (2022, pp. 106-110) appear to think, that there is some degree of arbitrariness involved in this modelling practice, in the sense that the choice of a certain model is not univocally dictated by the intrinsic characteristics of the target. Rather, the determination of the boundaries of the cognitive system in question will plausibly be significantly influenced by our explanatory goals. Specifically, depending on the behaviours *a modeller* individuates, and on how these behaviours are individuated, different alternative mechanisms and generative models may be relevant in originating them. It seems then that it is in principle impossible for us to be sure that a given objective generative model will match the subjective generative model the cognitive system makes use of. Remember that the subjective generative model is, by definition, a probability distribution concerning the cognitive system's take on the way its sensory states are generated. But the subjective generative model need not, and typically does not, precisely capture the actual way in which such states originate from the environmental influence upon the system. In other words, the subjective generative model is *not* the same as the actual generative *process*. This is particularly problematic if one intends to recover the subjective generative model as opposed to reconstructing the generative process. The reason is that, as external observers, on the one hand we have access to the observable behaviours which are a function of the subjective generative model, and on the other we may have access to part of the generative *process.*

[54] 'Complex living beings are composed of layer upon layer of Markov blankets […]. Different explanatory purposes drive us to highlight some of these blankets (of blankets) at the expense of others. But no blanket or set of blankets is privileged in and of itself'.

Consequently, we may incorrectly parametrise an objective generative model based on the information coming from our access to the latter, which may not correspond to the actual parameters of the subjective generative model.

An illustrative analogy may be helpful. Consider the earlier example in which I reach for my mosquito repellent while sitting next to a pond on a summer night. An external observer watching the scene may notice a mosquito biting my arm. Thus, they would assume that it is this environmental influence that causes some surprising sensory state, which, in turn, ultimately leads me to reach for the mosquito repellent as an attempt to avoid future similar surprising states. However, it is possible that I did not notice the mosquito bite, and I intended to apply it on myself just because I am particularly fond of its smell. Indeed, the behavioural data the observer takes into consideration in reconstructing what is going on may be partial (by observing the behaviour a little longer, it may turn out that I wanted to check the expiry date of the repellent) or altogether incorrect (I was really trying to reach for something else). Be that as it may, it is not just the observable behaviour of the cognitive system in question that drives our construction of the objective model. A non-negligible role is played by our consideration of facts that we (correctly or incorrectly) take to be part of the generative process, which we assume to be relevant inputs for the subjective generative model's delivering a certain behavioural output.

In short, one can arguably maintain that our modelling practices in the context of the FEP are importantly influenced by our explanatory interests. Consequently, one may be inclined to pick a certain Markov blanket rather than another, because, with respect to certain behavioural data, conceiving of the relevant cognitive system in one way rather than another may seem more appropriate. Therefore, there seems to be reason to think that it is not possible to find a principled criterion to individuate cognitive systems: there is no way to draw these boundaries fully independently from any explanatory interest. At any given time, there is a multitude of potential Markov blankets one may pick to separate a cognitive system from its environment, and, consequently, to constrain the subsequent objective generative model one will construct.

Some have challenged this final conclusion, which is sometimes labelled (e.g. by Kirchhoff and Kiverstein (2021)) "*proliferation*". For instance, Hohwy (2016) has made a case for privileging the brain as the cognitive system of interest. But, even if one were

to accept his proposal, the issue would just be pushed further back, rather than solved. For, even granting that the brain is by default the cognitive system of interest, the threat of an ensuing slippery slope shrinking down the dimension of that cognitive system would emerge (see Anderson (2017)). What constitutes the outer layer of the brain, and how should we model the different parts of the brain? Should we take just the outermost layer of neurons to constitute the Markov blanket? Why not the next inner one? This is referred to as the "*shrinkage*" problem by Kirchhoff and Kiverstein (2021), who also proposed their solution to both *proliferation* and *shrinkage*. According to them, there is not a single, persisting Markov blanket that demarcates the boundaries of a cognitive system throughout its existence. Nonetheless, at any given time it is possible to determine which among the various alternative Markov blankets one should choose to demarcate the boundaries of the cognitive system in question: the right Markov blanket is the one which, in terms of average free-energy minimisation, best accounts for the continued existence of the relevant system for a target period of time.

For now, it is not important to settle the dispute over the potential plurality of Markov blankets associated with a cognitive system. What matters for the present purposes is an implicit assumption that underlies all the views I have sketched, namely that there is *one correct way* to demarcate the boundaries of a cognitive system, and, consequently, to construct the objective generative model. It does not matter whether the subjective generative model separated from the environment by a Markov blanket is always the same (and identifiable with the brain, as per Hohwy). Nor does it matter whether it is diachronically negotiable (as per Kirchhoff and Kiverstein), or whether different subjective generative models are to be constructed relative to different observable behaviours (Parr, Pezzulo and Friston (2022, p.56)). Once a target phenomenon is pinpointed, to construct an objective generative model one has to assume that there is a unique subjective generative model associated with that phenomenon. This, at last, is the reason I believe that FEP-modelling, i.e. the construction of objective generative models, needs to be construed in terms of isomorphism, rather than in terms of some weaker morphism. Although there may not be a correct way to individuate the target phenomenon, namely the active inferences a cognitive system engages in, the underlying assumption is that for each putative target phenomenon there is a single subjective generative model, which needs to be reconstructed by modellers. If the

morphism between objective and subjective generative models is weaker than an isomorphism, then the objective generative model would fail to be sufficiently similar to the subjective generative model so as to ensure that it corresponds to the unique subjective generative model associated with the phenomenon to be modelled, whichever that may be and however that may be individuated.

Before moving on, one potential worry needs to be addressed. So far, I have argued that any morphism short of being an isomorphism would be insufficient for the purpose of picking the right subjective generative model. But there is a sense in which isomorphisms themselves may not be fully adequate for the purpose. As it has been long well-known (for an early elaboration of this point, see McLendon (1955)), isomorphisms are not as strong as they may appear at first. Indeed there are two senses in which, even though an isomorphism can be established, an objective generative model may fail to be correctly related to the intended subjective generative model. First, there may be more than just one isomorphism holding among two structures. Second, a structure typically is isomorphic to more than just one other structure. I do not think that the former consideration is especially troublesome for present purposes. The latter, instead, casts some doubts over the real adequacy of isomorphism as the morphism grounding the representational link meant to exist between subjective and objective generative models. In fact, it appears that isomorphisms are vulnerable to criticisms akin to the ones moved against less stringent morphisms: they are not strong enough to guarantee that the objective generative models will map onto the right subjective generative models.

I acknowledge the legitimacy of this worry. However, I believe that the sense in which isomorphisms are weaker than it would be desirable for the present purposes is different from the sense in which other morphisms are. Let me illustrate what I intend by this with a brief thought experiment.

Imagine that you take a perfectly clear picture (call it *pic1*) of a woman named Alice. It turns out that, unbeknownst to you, the woman you took a picture of is not really Alice, but her identical twin sister, Beth. Despite your photograph not being *really* a picture of Alice, it would make no difference in any relevant sense that this is so: it possesses all the features you can possibly be interested in, were you to examine the picture to learn something about the physical appearance of Alice. In this sense, it checks all the *required* boxes for being accepted as a representation of Alice, although that may

not be *enough* for really being such. Nonetheless, had the woman in the picture really been Alice, and not Beth, the picture would be a perfect representation of Alice: the reason why *pic1* is not perfectly adequate does not have to do with the properties of the picture itself, but on external circumstances. On the other hand, if your picture (*pic2*) also does not come out as perfectly as you hoped (say, her left arm is left out of the frame, or it is not clear whether she has freckles or not), the resulting picture may not be good enough to be used to learn everything you may be interested in about Alice's physical appearance. This would be the case even if Alice had been the subject of *pic2*: at least part of what makes *pic2* inadequate has to do with the properties of *pic2* itself.

The point I am driving at is that *pic1* and *pic2* are inadequate in different senses. Neither has *everything* it takes to be a perfectly useful representation of Alice (they are both insufficient). But in the case of *pic1*, this does not have to do with the features of the picture itself, and there is no practical difference as a result. On the contrary, in the case of *pic2*, because of some features of *pic2* itself, you may either be unable to learn certain things about Alice, or you may learn the wrong things (perhaps because, due to some light trick, her eyes appear of a different colour). This, I maintain, is analogous to the different ways in which isomorphisms and weaker morphisms are not perfectly up to the task when it comes to the representational relation meant to occur between subjective and generative models. Isomorphisms may not be sufficient to guarantee that a given structure is a univocal representation of some other structure, but this does not have to do with any of the features of the two structures *per se*. On the other hand, weaker morphisms are not sufficient because the structures connected by such functions are not suitable for the modelling practice in question. In terms of generative models, it seems that subjective generative models are such that they can be adequately captured by the objective generative models only if there is a function at least as strong as isomorphism in place, although the existence of such function may not be all is needed overall.

To conclude, I wish to reiterate that this is not to say that creating FEP-models requires structuralism to be globally true as an account of scientific representation, let alone that all scientific representations need to be isomorphic to their targets. Nonetheless, the assumption made in debates over the modelling practice involved in the FEP appears to be that the relation between objective and subjective generative models needs to be conceived in isomorphism-based structuralist terms. The generative models FEP-

scientists construct must be isomorphic to the generative models entailed by the free-energy minimising strategies adopted by cognitive systems. This is a necessary, although in all likelihood not sufficient, condition.

3. Against the realist stance

Let us take stock. In the first section I have presented the Free-Energy Principle, according to which living organisms manage to stay far from thermodynamical equilibrium by engaging in active inference, i.e. by engaging in strategies apt to minimise free-energy, a proxy for surprise. The separation of a given system from its environment, as well as the system's probabilistic "beliefs" about the way its sensory states are generated, are what FEP-models are meant to capture. More exactly, according to the FEP, the free-energy minimising strategies adopted by the relevant systems are the manifestation of a subjective generative model. It is then these subjective generative models that modellers try to reconstruct by elaborating their (objective) generative models, which should be understood as purporting to be isomorphic to the subjective generative models. It is now time to examine more closely how talk of "embodying a generative model" and "being delimited by a Markov blanket" are to be interpreted. This is what this section sets out to do, by applying the points raised in the second section to the particular case of the modelling practice guided by the Free-Energy Principle. The picture that will emerge discourages adopting a realist stance on FEP-models, because of the issues related to isomorphism-based representational processes.

As I mentioned earlier, it is no secret that isomorphism-based structuralism struggles to deal with misrepresentation. Broadly speaking, insofar as representations in general, and scientific models in particular, purport to represent their targets by containing information about their targets' features, they can end up misrepresenting their targets in three ways. First, they may fail to include some more or less important features of the target system, in which case they would be incomplete representations; this is the case, say, of a scale model of a building, which, differently from the real building, may not have windows, or a detailed inside. Second, they may contain information that does not correspond to actual properties possessed by the target system; this is the case of a planisphere, which features fictional lines indicating parallels and meridians. Third, they

may misrepresent the target system because of a combination of the first two ways to misrepresent; for instance, a toy model of the solar system may leave out features of the real system such as the presence of moons around Jupiter, while it might have thin metallic sticks keeping the planets suspended at fixed distances from each other (which obviously do not correspond to anything in the real solar system).

Misrepresentation is not inconsistent with partial representation in all three cases. The tiny building still represents the real building even if it is lacking some details (it is incomplete), the planisphere still represents the world even if it represents non-existing lines (it is inaccurate), and the toy solar system still represents the solar system even if it does not represent Jupiter's moons and even if the real planets are not kept in place by sticks (it is incomplete and inaccurate).

Nonetheless, it may happen that the representational process fails entirely because of any of the three cases presented. Imagine, as an illustrative analogy, that you ask me to draw a dromedary, and I draw a camel with two humps on its back. Dromedaries are camels with only one hump, so, in virtue of having drawn a camel with one hump too many, I have just drawn a generic camel, but I failed to represent a dromedary. Or, conversely, imagine that I am asked to draw a unicorn, and I draw what seems to be a normal horse. In virtue of lacking a crucial feature, my drawing fails to represent a unicorn. In both cases, the intuitions intended to be elicited are to the effect that I end up representing something else than what I intended to represent. I am not representing my targets at all, even if my representation is meant to represent them, and even if the first drawing is a complete and almost entirely accurate representation, while the second is an accurate and nearly complete representation.[55]

Now, if, as I have argued, FEP-modelling is based on isomorphism, it faces this sort of problem. If being isomorphic to its target is a necessary condition for an objective generative model to represent the relevant subjective generative model, then failure to establish an isomorphism between a model and its target amounts to the objective model's failure to represent its intended target. Injectivity without surjectivity (i.e. accuracy without completeness), surjectivity without injectivity (i.e. completeness without accuracy), bijectivity (i.e. injectivity and surjectivity together) without "relation-

[55] To reiterate, this analogy is not meant to be an example of how things really work in the case of drawings. Rather, it is meant to illustrate what the breakdown of the representational link due to misrepresentation would look like.

preserving-ness", or "relation-preserving-ness" without bijectivity; none of these options will do. Each of these four ways in which a function may fail to be an isomorphism between objective and subjective generative models, and which may result, in turn, in one of the three aforementioned ways in which misrepresentation might occur, is enough for the representational process to fail entirely, as far as isomorphism is concerned. This issue has important consequences with respect to the stance we should adopt on models based on the FEP. That is, if, for any given data set obtained from the observation of a cognitive system's behaviours there only corresponds one subjective generative model, and failure to establish an isomorphism between that model and the modellers' objective one leads to failure to represent the former, then we cannot be realist about the content of our objective generative models.

Let me elaborate. My criticism of realism with respect to FEP-models moves along different, weaker lines than other extant views. For instance, because of an observed systematic ambiguity between subjective and objective generative models, van Es (2021) denies that we should be realist about subjective generative models, in that reflections upon their objective counterparts do not warrant their reification. This is similar to, but subtly different from, what I maintain, as I do not take the distinction between subjective and objective generative models to be blurred and thus unable to warrant a realist stance on the former. What I do maintain is that, while the distinction may still be a meaningful one, we should not take our own (objective) models of such (subjective) models to perform their intended representational function. This is because the link between subjective and objective generative models is severed by the overwhelming likelihood that the required isomorphisms backing it up fail to obtain. In other words, I am not questioning the in-principle viability of realism, but only the actual effectiveness of the means by which such realism is meant to be bolstered. What I take issue with is not the possibility of making a realist move based on the sharpness of the distinction between subjective and objective models. I am, for the sake of the argument, granting that there are sufficiently solid conceptual grounds for this distinction, so much so that realism is not precluded. However, the realist move fails nonetheless, because the representational link which it requires breaks down.

In short, I am not ruling out in principle that the FEP may still ultimately turn out to get things right about how life and cognition work in general, including the fact that

cognitive systems' behaviours really do entail, in a realist and theorist-independent sense, (subjective) generative models. What I am denying is that this potentiality is enough to warrant taking our own (current) models to represent what really goes on when cognitive processes unfold. Hence, instrumentalism. I will return to this point in a moment.

Perhaps many will find this line of reasoning a little unusual. Generally speaking, antirealist arguments tend to deny that our theories "get things right" as opposed to just being empirically adequate, because the existence of some specified connection between theory-informed models and the target phenomena is necessary but not sufficient to warrant realism about the core claims of the theory in question. What I am claiming here proceeds in the opposite direction. Even if it really were the case that we are ultimately right about what the theory generally says concerning the kind of target phenomena, that would not be sufficient to take our theory-informed models to entertain the required sort of link with their target phenomena. To use an everyday analogy[56]: even if there is a cat on my bed, not every perceptual experience of a cat on my bed would thereby be veridical, as it may be a hallucination. Consequently, even if true, the realist belief that there is a cat on my bed would not be warranted. Similarly, even if it were not ultimately wrong to be realist about subjective generative models, this would not entail that the subjective generative models isomorphic to the corresponding objective generative models are the ones that we should be realist about. Consequently, even if true, the general claims afforded by the theory should not be construed in a realist way, as that would not be warranted.

It is important to point out that this, however, does not make the FEP a hopelessly empirically empty mathematical framework (for some critical discussions: Colombo and Wright (2021); Andrews (2021)), or at least not completely. If a theory can be said to have empirical content insofar as the claims it affords apply to the world because of the possibility of constructing models of the relevant phenomena, then the FEP, once coupled with some process theory, can meet this requirement in two cases. First, in case we overcome the difficulties undermining the isomorphism meant to hold between an objective generative model and its subjective counterpart. This is obviously a virtually impossible task for our scientific community, as it would require finding an objective,

[56] Again, as in the dromedary case, this is just meant to illustrate the point I am making. Many epistemologists would object that the possibility of hallucination does not typically undermine our perceptually justified beliefs.

ideal way to conceptualise cognitive systems, so as to solve the previously discussed issues related to drawing Markov blankets, for instance.

The second and significantly more viable possibility consists in embracing instrumentalism. That is, it consists in accepting that the relation between objective and subjective generative models is different from what it is currently thought to be. Under an instrumentalist reading of the modelling practices carried within the FEP framework, target systems may or may not actually engage in active inference. In either case, what warrants the modelling of the relevant target systems and phenomena by means of the FEP's array of conceptual tools is not the fact that objective generative models (and Markov blankets) correspond to the way their targets are. Rather, it is the fact that they are empirically adequate, in the sense of being explanatorily, descriptively, and predictively effective, at least to some degree.

The crucial point is that, in an instrumentalist setup, the representational link between objective and subjective generative models cannot be problematically severed, because such link is not established in the first place. This is because there is no commitment to the reality of subjective generative models. Subjective generative models are not the actual aspect of target systems objective generative models are meant to capture. Instead, they are a fictional (instrumental) conceptualisation of the target system, playing a role subordinate to empirical adequacy. That is, objective generative models aim at being empirically adequate models of their target systems, and this agenda is facilitated by their approximating the subjective generative models stemming from a construal of the target systems "as if" they were engaging in active inference. But, to reiterate, the failure of an objective generative model to be isomorphic to its subjective counterpart is not an issue. For, this does not amount to failing to be a model of the real target system.

One may argue that my understanding of realism commits what Kirchhoff, Kiverstein and Robertson (2022) have labelled the "literalist fallacy". According to Kirchhoff and colleagues, realist approaches to the FEP have been misguidedly criticised based on an overly demanding understanding of what realism maintains. In their view (which is akin to the one upheld by Godfrey-Smith (2003, 2009)), we should be realist about FEP-models as generalised models[57], that is, as capturing a "family" of phenomena.

[57] This terminology is originally introduced by Weisberg (2013).

Following Weisberg (2006, 2007), Kirchhoff and colleagues maintain that this understanding is immune to the sort of issues related to misrepresentation (in the sense I have been using the term here), insofar as the ultimate goal is to expunge them from our modelling practices. To claim otherwise, and accordingly criticise realism and uphold instrumentalism (as I did) is to commit the literalist fallacy; that is, it is to take realism to depend on the ability of our current models to be perfect, literal representations of their targets. This, in their view, is a mistake. Realism is not committed to such an unrealistic (pun intended) claim. Rather, realism is to be understood as maintaining that while our theories, *broadly speaking*, get things right about their targets, they will get the *details* right only in the long run, by eliminating, or reducing as much as possible, the use of idealisations.

As I understand this view, the gist is that FEP-models are *not literally false*[58]; rather, they are *approximately true*, and their presently being only approximately true does not constitute a problem for long-term realism. Simultaneously, this undermines attempts at putting forward instrumentalist readings of the FEP, as stronger reasons than current inaccuracy are required to bolster instrumentalism. However, although a full discussion of the form of realism they endorse is beyond the scope of the present discussion, some remarks in response to Kirchhoff's and colleagues' view need to be made.

First, it seems to me that their argument can be turned on its head. As mentioned above, instrumentalism differs from other forms of antirealism as it does not make the positive claim that scientific theories (and the FEP as such) *are not* true. They may or may not[59] be. What instrumentalism says is that, in the absence of irrefutable reasons to maintain their truth, scientific theories should be assessed purely in terms of their empirical (phenomenal) adequacy, without any commitment as to whether they "get things right", or even approximately right, with respect to what their targets really are like. In particular, the way I have argued for instrumentalism does not hinge on the falsity of current (and, likely, future) FEP-models. Rather, it hinges on their failure to establish the desired representational link with their targets. As a consequence, my case for instrumentalism should not be read as a manifestation of impatience, so to say, and unwillingness to wait for better, more precise models. Indeed, for a theory to be even just

[58] *Contra* Klein (2018, pp. 2253-54).
[59] In Van Es's and Hipólito's (2020, p.16) words: 'instrumentalism in itself is characterized by ontological agnosticism with regards to what actually makes a system tick'.

approximately true, as realists claim, that theory needs to represent its objects in the first place. Hence, if what I have been arguing for so far is correct, what realists need to do to block my instrumentalist argument is showing that there is a way for FEP-models to represent their intended targets in a way that accommodates misrepresentation. But, given what I have argued in section 2.2, it seems to me that the most plausible account of representation applying to the link between the objective and subjective generative models, in a realist context, is the isomorphism-based one I have been discussing.

In summary, by abandoning a realist approach to models based on the FEP, and adopting an instrumentalist one as an alternative, it is possible to avoid the problems associated with taking the representational relation between objective and subjective generative models to be based on isomorphism. In the next, final section I will consider how this position would reflect on the usefulness of the FEP as a theory from which to derive precise accounts of cognition specifically, i.e. marks of the cognitive.

4. Implications for marks of the cognitive

In the light of the argument presented in the previous section, it seems that the models of cognitive systems based on the FEP should not be construed in a realist way. This is because, I maintain, they would plausibly be meant to be isomorphic to the generative models (and Markov blankets) they describe. Hence, given that approximations, idealisations, and in general variably arbitrary design choices are virtually impossible, and indeed undesirable, to expunge from modelling practice, the required isomorphisms would systematically fail to obtain. As a consequence, even though scientific representation in general may not depend on isomorphisms to succeed, the representational attempts fail to go through in the case of FEP-models. In other words, insofar as FEP-models are understood in a realist way, they cannot succeed in representing their targets. This leaves only one option to FEP-theorists, namely conceiving of their models in an instrumentalist way. This, as we will see shortly, has interesting consequences for attempts at formulating a MOC in terms of the Free-Energy Principle.

4.1 The Free-Energy Principle is not a theory of cognition

One aspect of the FEP that the community of scholars variously concerned with it has acquired awareness of in very recent years is the fact that the FEP as such does not constitute a theory of cognition. Or, at least, it does not constitute a theory of cognition *specifically*. In fact, the claim that adaptive systems engage in active inference, thus increasing their likelihood of entering in unsurprising states and consequently remaining far from thermodynamical equilibrium (i.e. from death) does not specify anything that may distinguish cognition from other phenomena such as digestion, or even life itself[60]. While this may seem at first an interesting characteristic of the FEP, as an important connection between life and cognition is established[61], such a connection may lead, in some philosophical areas, to conceptual problems. If to be alive just is to engage in active inference, and if to be a cognitive system just is to be a system engaging in active inference, then there is no difference between being a cognitive system and being a living organism. But, given the current state of our knowledge and our conception of the natural world, the distinction between cognition and life is one that, at least for some philosophical purposes, ought to be preserved. This, of course, does not mean that the FEP as such needs to accommodate this distinction. Rather, it means that, if one intends to apply the FEP's conceptual framework to tackle philosophical issues that specifically have to do with cognition as a distinct natural phenomenon, then one needs to supplement it, at the level of the process theories paired with it, with some cognition-specific elements.

In short, one reason why the distinction between life and cognition is important is the interest that many philosophers of mind and the cognitive sciences have in the nature of cognition as a specific natural phenomenon. Indeed, one of the effects of the development of the 4E views on cognition (embodied, embedded, extended and enactive cognition) has been the increased felt need of a clear account of what cognition is. This need, clearly voiced for example by Adams (2010) and Wheeler (2019) (but for a sceptical take see Clark (2019)), has led to the formulation of a number of proposals. Some

[60] The worry that the scope of the FEP may end up being too broad may arise here. Such worry has been addressed by Kirchhoff et al. (2018), and by van Es and Kirchhoff (2021), by drawing a distinction between "mere" and "adaptive" active inference.

[61] Kirchhoff, Froese (2017); Kirchhoff (2018a); Bruineberg, Kiverstein and Rietveld (2018). This idea appears also in the literature on autopoiesis (Maturana and Varela (1980)) and on enactivism (Thompson, (2007); Di Paolo, (2009))

proposals (Adams and Aizawa (2008); Rowlands (2010); Adams and Garrison (2013)) are directly linked to the debate over extended cognition stemming from Clark and Chalmers (1998), while others are more closely related to the literature over the contrast between anthropogenic and biogenic approaches to cognition (Lyon (2006); Van Duijn, Keijzer and Franken (2006); Keijzer (2021)).

To obtain a theory specifically of cognition based on the FEP one needs therefore to add further constraints to the core claims constituting the FEP. Differently put, one needs to show how cognition enables adaptive systems to minimise their free-energy, while at the same time differentiating it from other free-energy minimising phenomena. This is what, for instance, Kiverstein and Sims do, arguing that cognition ought to be understood in terms of allostatic control, which is 'prospective behaviour directed at avoiding the anticipated divergence from homeostatic setpoints' (2021, p.25). Their criterion draws on the FEP: insofar as homeostasis is achieved by entering states associated with low free-energy, allostatic control results in free-energy minimisation. Moreover, what Kiverstein's and Sims's account stresses is the proactive nature of the behaviours involved in allostatic control. That is, such behaviours must not be purely reactive, but also, and crucially, anticipatory. This allows genuine instances of cognition to be discerned from other free-energy minimising processes. For instance, the circulatory system of some creature plausibly should not be considered responsible for any cognitive activity that creature may be said to perform, as its behaviours would be purely reactive, instead of proactive.

Now, I am not concerned here with the assessment of the strength of Kiverstein's and Sims's proposal specifically, nor of any other alternative account[62]. What I intend to point out is just that it is possible to elaborate an account of cognition in line with the FEP, as it has indeed been done. But in order to do so, one needs to specify some distinctive characteristics of cognition that other free-energy minimising strategies do not possess: cognition is just one of the ways in which adaptive systems stay attuned to their environment in a suitable way for their survival, but it is not the only one, and it may need to be distinguished from the others.

[62] Indeed, Kiverstein's and Sims's work is a response to Corcoran, Pezzulo and Hohwy (2020). Corcoran and colleagues also maintain that active inference is what grounds the appearance of cognitive phenomena, but, unlike Kiverstein and Sims, they spell their account in terms of counterfactual active inference, rather than in terms of allostatic control.

4.2 Instrumentalism and the mark of the cognitive

So, the FEP is not, *per se*, a theory of cognition, but it can be the conceptual framework in which more specific proposals for a mark of the cognitive are couched. However, based on the points raised throughout this chapter, the models of cognitive systems that might be constructed in such a conceptual framework encounter a series of problems if looked at from a realist perspective.

Realism is an appealing position, especially when it comes to the characterisation of natural phenomena like cognition. Being realists about what our theories of cognition have to say means taking our theories to capture, to different degrees of accuracy and completeness, what the nature of cognition is, as well as how cognitive systems are structured and work. But realism with respect to FEP-models is problematic. Consider again Kiverstein's and Sims's proposal that cognition has to be understood in terms of allostatic control. Adopting a realist stance on this view means underwriting the claim that cognitive systems display allostatic control independently from any external observer's acknowledgement of this fact, so that allostatic control is not just a theoretical construction helping us to make sense of what cognitive systems do. Furthermore, any model representing the (subjective) generative models supporting allostatic control would have to be taken as faithfully depicting the models really in play in the generation of these phenomena. However, as we saw, this risks being too optimistic a view, one which fails to take into consideration the unavoidable creeping in of theorist-dependent design choices in the elaboration of such models.

This can be noted by taking a closer look at one of the examples Kiverstein and Sims discuss, namely the case of slime mould, *Physarum polycephalum* (2021, pp. 19-20). Slime mould has been observed to slow down its motion, while foraging for food, when it anticipates the periodic occurrence of a dry stimulation which would normally elicit such behaviour (Saigusa, Tero, Nakagaki, and Kuramoto (2008)). This anticipatory behaviour motivates Kiverstein's and Sims's conclusion that slime mould manifests cognition, because, by engaging in this proactive, not purely reactive, behaviour, it reduces its expected free-energy (it decreases the likelihood of entering future surprising states).

Now, for the "allostatic control" mark of the cognitive to be one that stems from the FEP, it is not enough that putative cognitive systems display some form of allostatic

control: they also need to do so in accordance with the concepts the FEP is concerned with. This means that one needs to create a model of the system under examination by identifying its Markov blanket (its sensory and active states) and by specifying its generative model. But this is where the modelling issues I have been concerned with enter the scene. As the authors of the study on slime mould explain, the locomotion of the entire organism depends on multiple chemical oscillators (Saigusa, Tero, Nakagaki and Kuramoto (2008, p.3)), which means that the active states of the organism (those corresponding to its movement at different speeds) are determined by the internal states of the organism, to be understood in terms of such chemical oscillators. In turn, these internal states are influenced by other states the system enters in: the relevant periodic sensory states the slime mould enters in as a result of dry stimulation on the one hand, and other internal states on the other hand, which, over time and provided that the periodic dry stimulation is not offered in the meantime, "reset" the system to its baseline condition, so that the periodic slowing-down ceases.

All these factors need to be appropriately expressed in our FEP-informed model of the activities of the slime mould. If one fails to precisely model all the relevant chemical oscillators (an arguably difficult task), this may or may not have important consequences in terms of empirical adequacy of the resulting model, but it surely would be enough for the objective generative model in question not to be isomorphic to the subjective generative model. And this, as we have seen earlier, makes the representational relation between the two fall apart, thus warranting an instrumentalist, instead of a realist, take on the objective generative model. Such a model may be empirically adequate, but it does not capture what "really" is going on in the organism under examination, from the organism's perspective.

The case of the slime mould is a relatively simple one, and one may be understandably not too worried about interpreting a FEP-informed mark of the cognitive instrumentally. But the disputes that led many to think that an account of cognition is of crucial importance make instrumentalism undesirable. Consider the debate over the extended cognition view. Is Clark's and Chalmers's (1998) well-known imaginary Otto, a person affected by Alzheimer's disease who heavily relies on the information written in his notebook to navigate the world, involved in an extended cognitive system whose boundaries encompass not only Otto himself, but also his notebook? According to

Kirchhoff and Kiverstein (2019, 2021), yes, because taking this to be the case would allow us to have a better explanation for the continued existence of the "Otto + notebook" system than we would have otherwise. However, while this is consistent with an instrumentalist approach to the issue, instrumentalists of different inclinations may not feel the need to include the notebook in the picture. This is indeed an admissible move, as it would only require us to consider the states Otto himself enters in as a result of interacting with his notebook as sensory states rather than internal states statistically conditioned by other internal states. In terms of the empirical adequacy of our modelling, all else being equal, there would not be much of a difference. Otto would enter the same active states, and he would tend to minimise his free-energy in similar ways. Furthermore, proponents of the "Otto + notebook model" or of the "Otto model" would have an equally questionable claim to the correctness of their conception of the cognitive system in question. For, any matter of fact able to settle the dispute would be out of reach in virtue of the overall instrumentalist framework in which the disputants would be working. As long as their models are equally empirically adequate (provided that the contenders can even agree on how to assess this), neither disputant can expect the world to tip the scale in favour of their preferred model, because both FEP-informed ways of conceiving the cognitive system would ultimately fail to represent the real cognitive system.

To summarise, instrumentalism may not prevent FEP-based accounts of cognition from being informative about the nature of cognition. There is no question that, for instance, Kiverstein and Sims make an interesting and substantial claim about what cognition is (regardless of whether it is correct or not, and of whether there is an actual need to account for cognition specifically). Nonetheless, as I have tried to show just now, going instrumentalist would make some of the philosophical disputes on cognition, which led to the felt need for an account of cognition, impossible to be solved, even if a mark of the cognitive were offered.

Conclusion

In this chapter, I have introduced and discussed a popular conceptual framework that can serve as the background for the formulation of biogenic marks of the cognitive. In particular, I have argued that we should be instrumentalists, instead of realists, about

the models of the Free-Energy Principle. Instead of arguing directly for this position by questioning whether adaptive systems (and cognitive systems in particular) should be literally taken to engage in active inference, I made my case using some insights coming from the literature on scientific modelling.

After having introduced the FEP and the use of models involved in it, I have argued that objective generative models should be interpreted as intended to be isomorphic to subjective generative models, and that it is in virtue of this isomorphism that the former represent the latter. But, if the representational process requires the existence of an isomorphism between the representing and the represented structure, and since specifying the representing structure is largely a matter of more or less arbitrary design choices, then it appears that one can hardly ever hope that objective generative models will represent their intended subjective counterparts at all. If this is so, then there is no reason to maintain that FEP-models (models of adaptive systems based on the FEP) are to be interpreted in a realist way. In fact, for realism about FEP-models to be warranted, such models should at the very least be representations of their targets, but that is very likely not to be the case.

Finally, I have concluded with some reflections on the FEP as a source of accounts of cognition. The FEP *per se* is not specifically a theory of cognition, but attempts at formulating a mark of the cognitive based on the FEP can and have been made. However, since we should be instrumentalists about FEP-models, while a mark of the cognitive based on the FEP may further our understanding of cognition, it is unlikely to help us to settle in any specific case at least some of the philosophical disputes whose solution is thought to need an account of cognition.

Concluding Remarks

It is time to draw some conclusions and to make a few final remarks on the discussion about the issue of finding a mark of the cognitive which unfolded over the past six chapters. In the introduction, I have given some reasons why one may be interested in finding a MOC. It will thus be interesting to see where we are at in terms of achieving the related goals; but first, let me briefly summarise the general structure of the present work.

In Chapter 1, I presented and commented on three extant proposals for a MOC, namely Adams's and Aizawa's, Adams's and Garrison's, and Rowlands's ones. Drawing on the insights coming from the analysis of the notion of nonderived content developed in the first chapter, in Chapter 2 I began my argument in support of the 1PP condition on cognition. This was done in primarily epistemic terms, and, in Chapter 3, I further developed such arguments, this time in metaphysical terms: the existence of a 1PP associated with cognition is necessary not just for our understanding of cognition, but also for its very occurrence. Then, in Chapter 4, I started my criticism of the 1PP condition, showing how some of its implications go against scientific consensus. In Chapter 5 I have further strengthened my criticism, showing how the exclusion problem plagues not only the 1PP condition on cognition, but also anthropogenic approaches to cognition more generally (of which the 1PP proposal is an example). Thus, I recommended the adoption of a biogenic approach to the issue of finding a MOC, and in Chapter 6 I have extensively discussed the FEP framework, a well-developed mathematical and conceptual framework which can be and has been used to formulate biogenic proposals for a MOC.

As mentioned in the introduction of this thesis, in addition to being an interesting research project in its own right, the core reason why we should be interested in finding a MOC is that a MOC would put us in an optimal position to find the answers to several philosophical questions, thanks to the three main theoretical benefits that it would, or should, afford. First, a MOC would allow us to distinguish the domain of the cognitive from that of the non-cognitive. Second, it would (ideally) provide us with a criterion for drawing the boundaries of cognitive systems. Third, it would allow us (again, ideally) to understand what it means for a cognitive process to belong to a cognitive subject. Based

on the discussion from the last chapter, a MOC framed in terms of the FEP would be able to do all these things, but with one caveat. Elaborating a biogenic MOC against the FEP conceptual background would of course result in a demarcation of the domain of the cognitive from the domain of the noncognitive (this is, after all, what a MOC at bottom ought to do). Not only that, but it would also specify what it means for a cognitive process to belong to a cognitive subject: it is to minimise the free-energy associated with the system in question in a cognitive way (which is specified by the components of the particular MOC). Finally, it would provide us with an understanding of how the boundaries of a cognitive system are to be drawn; that is, the boundaries of a cognitive system are set via the Markov blankets formalism. However, if, as I have argued, we should be instrumentalists about the models of the FEP, the way a FEP-based MOC construes the boundaries of a cognitive system would not be practically useful for us, as we would not be able to actually individuate such boundaries for any particular cognitive system.

In this sense, then, a FEP-based MOC would provide us with what we need to answer many, although not all, of the questions that motivate the search for a MOC in the first place. Of course, this may be enough to be content, but one may wonder whether different ways to go down the biogenic path could deliver even better results. I have been extensively discussing the FEP because it is an undeniably well-developed mathematical framework and it is strikingly promising as a background against which to pursue a biogenic approach. With that being said, buying the FEP story is not the only way to begin one's pursuit of a biogenic approach. Nevertheless, as far as I am aware, there are not many equally promising competitors of the FEP, so it is reasonable to stick with it for the time being. Still, I am open to the possibility of formulating a MOC not in terms of the FEP, as long as this will be the product of a biogenic approach to the issue.

There is one more point worth bringing up in this concluding section. Throughout this thesis, I devoted at least as much energy to the defence of the 1PP condition on cognition as I did to its criticism. By now, it should be clear that engaging in such an extensive discussion was not just a dialectical exercise. Rather, it served the constructive purpose of laying solid foundations for my endorsement of the general biogenic approach to the issue of finding a MOC. By showing how what I take to be the most plausible proposal for a necessary condition for cognition is affected by a series of fatal issues,

some of which due to its being a product of an anthropogenic approach, I offered robust contrastive reasons to follow a biogenic approach. Furthermore, throughout the discussion I had the chance to explain at length what is at stake in the search for a MOC: the understanding of what cognition is via the formulation of a (set of) condition(s) for its occurrence, the demarcation of the boundaries of cognitive systems, and an explanation of what it means for a cognitive process to belong to a subject. For these reasons, even though I ended up dismissing the 1PP condition on cognition, my analysis of that condition was fruitful.

So, what comes next? Despite having been a hotly debated topic in recent years, the search for a MOC is not yet coming to an end. Therefore, the next step will consist in assessing the extant biogenic proposals, be they FEP-based or not, learning from their potential shortcomings (especially with respect to their ability to meet the expectations associated with a MOC), and moving forward towards the final account of cognition. I hope that the results of the present work will help shaping the future of the quest for a MOC.

References

1. Adams, F. (2010). Why we still need a mark of the cognitive. *Cognitive Systems Research*, 11(4), 324-331. https://doi.org/10.1016/j.cogsys.2010.03.001
2. Adams, F., & Aizawa, K. (2001). The bounds of cognition. *Philosophical psychology*, 14(1), 43-64. https://doi.org/10.1080/09515080120033571
3. Aizawa, K., & Adams, F. (2005). Defending non-derived content. *Philosophical Psychology*, 18(6), 661-669. https://doi.org/10.1080/09515080500355186
4. Adams, F., and Aizawa, K. (2008). *The Bounds of Cognition*. Oxford: Blackwell.
5. Adams, F., & Aizawa, K. (2010a). Defending the bounds of cognition. In R. Menary (Ed.), *The extended mind*, 67-80. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/9780262014038.003.0004
6. Adams, F., & Aizawa, K. (2010b). The value of cognitivism in thinking about extended cognition. *Phenomenology and the Cognitive Sciences*, 9(4), 579-603. https://doi.org/10.1007/s11097-010-9184-9
7. Adams, F., & Garrison, R. (2013). The mark of the cognitive. *Minds and Machines*, 23(3), 339-352. https://doi.org/10.1007/s11023-012-9291-1
8. American Academy of Sleep Medicine. *International Classification of Sleep Disorders*, 3rd ed. Darien, IL: American Academy of Sleep Medicine; 2014.
9. Allen, M., & Friston, K. J. (2018). From cognitivism to autopoiesis: towards a computational framework for the embodied mind. *Synthese*, *195*(6), 2459-2482. https://doi.org/10.1007/s11229-016-1288-5
10. Anderson, M. (2017). Of Bayes and bullets: An embodied, situated, targeting-based account of predictive processing. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing (Vol. 3*). MIND Group: Frankfurt am Main.
11. Andrews, M. (2021). The math is not the territory: navigating the free energy principle. *Biology & Philosophy*, 36(3), 1-19. https://doi.org/10.1007/s10539-021-09807-0

12. Avidan, A. Y., & Kaplish, N. (2010). The parasomnias: epidemiology, clinical features, and diagnostic approach. *Clinics in chest medicine*, 31(2), 353-370. https://doi.org/10.1016/j.ccm.2010.02.015
13. Barrett, A. B., & Seth, A. K. (2011). Practical measures of integrated information for time-series data. *PLoS computational biology*, 7(1), e1001052. https://doi.org/10.1371/journal.pcbi.1001052
14. Barrett, A. B. (2014). An integration of integrated information theory with fundamental physics. *Frontiers in psychology*, *5*, 63. https://doi.org/10.3389%2Ffpsyg.2014.00063
15. Barrett, A. B. (2016). A comment on Tononi & Koch (2015)'Consciousness: here, there and everywhere?'. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1687), 20140198. https://doi.org/10.1098/rstb.2014.0198
16. Barrett, A. B., & Mediano, P. A. (2019). The Phi measure of integrated information is not well-defined for general physical systems. *Journal of Consciousness Studies*, *26*(1-2), 11-20. https://doi.org/10.48550/arXiv.1902.04321
17. Bartels, A. (2006). Defending the structural concept of representation. *Theoria: An International Journal for Theory, History and Foundations of Science*, 21(1), 7-19. https://doi.org/10.1387/theoria.550
18. Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge University Press.
19. Bassetti, C., Vella, S., Donati, F., Wielepp, P., & Weder, B. (2000). SPECT during sleepwalking. *The Lancet*, 356(9228), 484-485. https://doi.org/10.1016/s0140-6736(00)02561-7
20. Bayne, T. (2018). On the axiomatic foundations of the integrated information theory of consciousness. *Neuroscience of consciousness*, *2018*(1). https://doi.org/10.1093/nc/niy007
21. Bechtel, W. 2008. *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. New York: Taylor and Francis.
22. Bennett, K. (2003). Why the exclusion problem seems intractable, and how, just maybe, to tract it. *Noûs*, 37(3), 471-497. https://doi.org/10.1111/1468-0068.00447

23. Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and brain sciences*, 18(2), 227-247. https://doi.org/10.1017/S0140525X00038188
24. Block, N. (2003). Do causal powers drain away?. *Philosophy and Phenomenological Research*, 67(1), 133-150. https://doi.org/10.1111/j.1933-1592.2003.tb00029.x
25. Boly, M., Perlbarg, V., Marrelec, G., Schabus, M., Laureys, S., Doyon, J., Pélégrini-Issac, M., Maquet, P., & Benali, H. (2012). Hierarchical clustering of brain activity during human nonrapid eye movement sleep. *Proceedings of the National Academy of Sciences*, 109(15), 5856-5861. https://doi.org/10.1073%2Fpnas.1111133109
26. Broughton, R. J. (1968). Sleep Disorders: Disorders of Arousal?: Enuresis, somnambulism, and nightmares occur in confusional states of arousal, not in" dreaming sleep.". *Science*, 159(3819), 1070-1078. https://doi.org/10.1126/science.159.3819.1070
27. Bruineberg, J., Dołęga, K., Dewhurst, J., & Baltieri, M. (2022). The Emperor's New Markov Blankets. *Behavioral and Brain Sciences*, *45*, E183. https://doi.org/10.1017/S0140525X21002351
28. Bruineberg, J., Kiverstein, J., & Rietveld, E. (2018). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese*, 195(6), 2417–2444. https://doi.org/10.1007/s11229-016-1239-1
29. Bueno, O. (1997). Empirical adequacy: A partial structures approach. *Studies in History and Philosophy of Science Part A*, 28(4), 585-610. https://doi.org/10.1016/S0039-3681(97)00012-5
30. Bueno, O., & French, S. (2011). How Theories Represent. *British Journal for the Philosophy of Science*, 62(4), 857-894. https://doi.org/10.1093/bjps/axr010
31. Buller, D. J. (2006). *Adapting minds: Evolutionary psychology and the persistent quest for human nature*. Cambridge, MA: MIT press.
32. Callender, C., & Cohen, J. (2006). There is no special problem about scientific representation. *Theoria: An International Journal for Theory, History and Foundations of Science*, 21(1), 67-85. https://doi.org/10.1387/theoria.554

33. Castelnovo, A., Riedner, B. A., Smith, R. F., Tononi, G., Boly, M., & Benca, R. M. (2016). Scalp and source power topography in sleepwalking and sleep terrors: a high-density EEG study. *Sleep*, 39(10), 1815-1825. https://doi.org/10.5665/sleep.6162

34. Cavallero, C., Cicogna, P., Natale, V., Occhionero, M., & Zito, A. (1992). Slow wave sleep dreaming. *Sleep*, 15(6), 562-566. https://doi.org/10.1093/sleep/15.6.562

35. Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7-19. https://doi.org/10.1093/analys/58.1.7

36. Clark, A. (2005). Intrinsic content, active memory and the extended mind. *Analysis*, 65(1), 111. https://doi.org/10.1093/analys/65.1.1

37. Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind.* New York: Oxford University Press.

38. Clark, A. (2017). How to knit your own Markov blanket. In T. Metzinger and W. Wiese (Eds), *Philosophy and Predictive Processing (Vol. 3)*. MIND Group: Frankfurt am Main.

39. Clark A (2019) Replies to critics: in search of the embodied, extended, enactive, predictive (EEE-P) mind. In: Colombo M, Irvine E, Stapleton M (eds) *Andy Clark and his Critics*. Oxford University Press, pp 266–303.

40. Clavel Vázquez, M. J., & Wheeler, M. (2018). Minding Nature: Gallagher and the Relevance of Phenomenology to Cognitive Science. *Australasian Philosophical Review*, *2*(2), 145-158. https://doi.org/10.1080/24740500.2018.1552085

41. Colombo, M., Elkin, L., & Hartmann, S. (2021). Being realist about Bayes, and the predictive processing theory of mind. *The British Journal for the Philosophy of Science*, 72(1), 185-220. https://doi.org/10.1093/bjps/axy059

42. Colombo, M., & Wright, C. (2021). First principles in the life sciences: the free-energy principle, organicism, and mechanism. *Synthese*, 198(14), 3463-3488. https://doi.org/10.1007/s11229-018-01932-w

43. Conant, R. C., & Ross Ashby, W. (1970). Every good regulator of a system must be a model of that system. *International journal of systems science*, 1(2), 89-97. https://doi.org/10.1080/00207727008920220

44. Corcoran, A. W., Pezzulo, G., & Hohwy, J. (2020). From allostatic agents to counterfactual cognisers: active inference, biological regulation, and the origins of cognition. *Biology & Philosophy*, *35*(3), 1-45. https://doi.org/10.1007/s10539-020-09746-2

45. Cummins, R. (1996*). Representations, targets, and attitudes*. Cambridge, MA: MIT Press.

46. D'Ambrosio, S., Castelnovo, A., Guglielmi, O., Nobili, L., Sarasso, S., & Garbarino, S. (2019). Sleepiness as a local phenomenon. *Frontiers in Neuroscience*, 13, 1086. https://doi.org/10.3389/fnins.2019.01086

47. Dell, P. F. (2006). A new model of dissociative identity disorder. *Psychiatric Clinics*, *29*(1), 1-26. https://doi.org/10.1016/j.psc.2005.10.013

48. Dennett, D.C. (1988) Quining qualia, in Marcel, A.J. & Bisiach, E. (eds.) *Consciousness in Modern Science*, 42–77. Oxford: Oxford University Press

49. Dennett, D. C. (1991). *Consciousness explained*. Boston: Little, Brown.

50. Dennett, D. C. (1994). The myth of original intentionality. In Dietrich, E. (Ed.). *Thinking computers and virtual persons: Essays on the intentionality of machines*. Academic Press.

51. Dennett, D. C. (2001). *The fantasy of first-person science*.

52. Dennett, D. C. (2007). Heterophenomenology reconsidered. *Phenomenology and the Cognitive Sciences*, 6(1-2), 247-270. https://doi.org/10.1007/s11097-006-9044-9

53. Desjardins, M. È., Carrier, J., Lina, J. M., Fortin, M., Gosselin, N., Montplaisir, J., & Zadra, A. (2017). EEG functional connectivity prior to sleepwalking: evidence of interplay between sleep and wakefulness. *Sleep*, 40(4), zsx024. https://doi.org/10.1093/sleep/zsx024

54. Di Paolo, E. (2009). Extended Life. *Topoi*, 28, 9–21. https://doi.org/10.1007/s11245-008-9042-3

55. Di Paolo, E., Thompson, E., & Beer, R. (2022). Laying down a forking path: Tensions between enaction and the free energy principle. *Philosophy and the Mind Sciences*, *3*. https://doi.org/10.33735/phimisci.2022.9187

56. Doerig, A., Schurger, A., & Herzog, M. H. (2021). Hard criteria for empirical theories of consciousness. *Cognitive neuroscience*, 12(2), 41-62. https://doi.org/10.1080/17588928.2020.1772214
57. Dretske, F. (1981). *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.
58. Dretske, F. (1988). *Explaining behavior*. Cambridge, MA: MIT Press.
59. Facchin, M. (2021). Extended predictive minds: do Markov Blankets matter?. *Review of Philosophy and Psychology*, 1-30. https://doi.org/10.1007/s13164-021-00607-9
60. Flanagan, O. (1995). Consciousness and the natural method. *Neuropsychologia*, 33(9), 1103-1115. https://doi.org/10.1016/0028-3932(95)00051-4
61. Fodor, J. (1987). *Psychosemantics*. Cambridge, MA: MIT Press.
62. Fodor, J. (1990). *A theory of content and other essays*. Cambridge, MA: MIT Press.
63. Frankish, K. (2016). Illusionism as a theory of consciousness. *Journal of Consciousness Studies*, 23(11-12), 11-39.
64. Friston, K. (2009). The free-energy principle: a rough guide to the brain?. *Trends in cognitive sciences*, 13(7), 293-301. https://doi.org/10.1016/j.tics.2009.04.005
65. Friston, K. (2010). The free-energy principle: a unified brain theory?. *Nature reviews neuroscience*, 11(2), 127-138. https://doi.org/10.1038/nrn2787
66. Friston, K. (2012). A free energy principle for biological systems. *Entropy*, 14(11), 2100-2121. https://doi.org/10.3390/e14112100
67. Friston, K. (2013). Life as we know it. *Journal of the Royal Society Interface*, *10*(86), 20130475. https://doi.org/10.1098/rsif.2013.0475
68. Friston, K., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of physiology-Paris*, 100(1-3), 70-87. https://doi.org/10.1016/j.jphysparis.2006.10.001
69. Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese*, *193*(2), 559-582. https://doi.org/10.1007/s11229-015-0762-9
70. Godfrey-Smith, P. (2003). *Theory and Reality: An Introduction to the Philosophy of Science*. Chicago: University of Chicago Press.

71. Godfrey-Smith, P. (2009). Models and Fictions in Science. *Philosophical Studies*, 143(1), 101–116. https://doi.org/10.1007/s11098-008-9313-2
72. Guilleminault, C., Poyares, D., Abat, F., & Palombini, L. (2001). Sleep and wakefulness in somnambulism: a spectral analysis study. *Journal of psychosomatic research*, 51(2), 411-416. https://doi.org/10.1016/s0022-3999(01)00187-8
73. Hohwy, J. (2013). *The predictive mind*. Oxford: Oxford University Press.
74. Hohwy, J. (2015). The neural organ explains the mind. In T. Metzinger & J. M. Windt (Eds). *Open MIND*. 19(T). Frankfurt am Main: MIND group. https://doi.org/10.15502/9783958570016
75. Hohwy, J. (2016). The self-evidencing brain. *Noûs*, 50(2), 259-285. https://doi.org/10.1111/nous.12062
76. Hohwy, J. (2017). How to entrain your evil demon. In T. Metzinger and W. Wiese (Eds), *Philosophy and Predictive Processing (Vol. 3)*. MIND Group: Frankfurt am Main.
77. Hohwy, J. (2021). Self-supervision, normativity and the free energy principle. *Synthese*, *199*(1), 29-53. https://doi.org/10.1007/s11229-020-02622-2
78. Howell, M. J. (2012). Parasomnias: an updated review. *Neurotherapeutics*, 9(4), 753-775. https://doi.org/10.1007/s13311-012-0143-8
79. Kahn, D., Pace-Schott, E. F., & Hobson, J. A. (1997). Consciousness in waking and dreaming: the roles of neuronal oscillation and neuromodulation in determining similarities and differences. *Neuroscience*, 78(1), 13-38. https://doi.org/10.1016/S0306-4522(96)00550-7
80. Kallestrup, J. (2006). The causal exclusion argument. *Philosophical Studies*, *131*(2), 459-485. https://doi.org/10.1007/s11098-005-1439-x
81. Keijzer, F. (2021). Demarcating cognition: the cognitive life sciences. *Synthese*, 198(1), 137-157. https://doi.org/10.1007/s11229-020-02797-8
82. Kim, J. (1997). Does the problem of mental causation generalize?. In *Proceedings of the Aristotelian Society* (97), 281-297. Aristotelian Society, Wiley. https://doi.org/10.1111/1467-9264.00017
83. Kim, J. 1998. *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. Cambridge, MA: MIT Press.

84. Kim, J. 2007. *Physicalism, or Something Near Enough*. Princeton, NJ: Princeton University Press.
85. Kirchhoff, M. D. (2018). Autopoiesis, free energy, and the life–mind continuity thesis. *Synthese*, *195*(6), 2519-2540. https://doi.org/10.1007/s11229-016-1100-6
86. Kirchhoff, M. D., & Froese, T. (2017). Where there is life there is mind: In support of a strong life-mind continuity thesis. *Entropy*, *19*(4), 169. https://doi.org/10.3390/e19040169
87. Kirchhoff, M. D., & Kiverstein, J. (2019*). Extended consciousness and predictive processing: A third-wave view*. Routledge.
88. Kirchhoff, M. D., & Kiverstein, J. (2021). How to determine the boundaries of the mind: A Markov blanket proposal. *Synthese*, 198(5), 4791-4810. https://doi.org/10.1007/s11229-019-02370-y
89. Kirchhoff, M., Kiverstein, J., & Robertson, I. (2022). The literalist fallacy and the free energy principle: On model-building, scientific realism and instrumentalism. *Br. J. Philos. Sci*. https://doi.org/10.1086/720861
90. Kirchhoff, M., Parr, T., Palacios, E., Friston, K., & Kiverstein, J. (2018). The Markov blankets of life: autonomy, active inference and the free energy principle. *Journal of The Royal Society Interface*, 15(138), 20170792. https://doi.org/10.1098/rsif.2017.0792
91. Kiverstein, J., & Sims, M. (2021). Is free-energy minimisation the mark of the cognitive?. *Biology & Philosophy*, 36(2), 1-27. https://doi.org/10.1007/s10539-021-09788-0
92. Klein, C. (2018). What do predictive coders want?. *Synthese*, 195, 2541–2557. https://doi.org/10.1007/s11229-016-1250-6
93. Koch, C. (2019). *The feeling of life itself: Why consciousness is widespread but can't be computed*. Cambridge, MA: MIT Press.
94. Kriegel, U. (2011). *The sources of intentionality*. Oxford University Press.
95. Krueger, J. M., Nguyen, J. T., Dykstra-Aiello, C. J., & Taishi, P. (2019). Local sleep. *Sleep medicine reviews*, 43, 14-21. https://doi.org/10.1016/j.smrv.2018.10.001

96. List, C., & Menzies, P. (2009). Nonreductive physicalism and the limits of the exclusion principle. *The Journal of Philosophy*, *106*(9), 475-502. https://doi.org/10.5840/jphil2009106936

97. Lowe, E. J. (2000). Causal closure principles and emergentism. *Philosophy*, *75*(294), 571-585. https://doi.org/10.1017/S003181910000067X

98. Lyon, P. (2006). The biogenic approach to cognition. *Cognitive Processing*, 7(1), 11-29. https://doi.org/10.1007/s10339-005-0016-8

99. Massimini, M., Ferrarelli, F., Huber, R., Esser, S. K., Singh, H., & Tononi, G. (2005). Breakdown of cortical effective connectivity during sleep. *Science*, 309(5744), 2228-2232. https://doi.org/10.1126/science.1117256

100. Massimini, M., Ferrarelli, F., Murphy, M. J., Huber, R., Riedner, B. A., Casarotto, S., & Tononi, G. (2010). Cortical reactivity and effective connectivity during REM sleep in humans. *Cognitive neuroscience*, 1(3), 176-183. https://doi.org/10.1080/17588921003731578

101. Maturana, H., & Varela, F. (1980). *Autopoiesis and cognition: The realization of the living*. Dordrecht: D. Reidel.

102. McLendon, H. J. (1955). Uses of Similarity of Structure in Contemporary Philosophy. *Mind*, 64(253), 79–95. http://www.jstor.org/stable/2251045

103. Mediano, P. A. M., Rosas, F., Carhart-Harris, R. L., Seth, A. K., & Barrett, A. B. (2019a). Beyond integrated information: A taxonomy of information dynamics phenomena. *ArXiv190902297 Phys. Q-Bio*. Available at: https://doi.org/10.48550/arXiv.1909.02297

104. Mediano, P. A., Seth, A. K., & Barrett, A. B. (2019). Measuring integrated information: Comparison of candidate measures in theory and simulation. *Entropy*, 21(1), 17. https://doi.org/10.3390/e21010017

105. Menary, R. (2006). Attacking the bounds of cognition. *Philosophical Psychology*, 19(3), 329344. https://doi.org/10.1080/09515080600690557

106. Menary, R. (2010). The holy grail of cognitivism: a response to Adams and Aizawa. *Phenomenology and the Cognitive Sciences*, 9(4), 605-618. https://doi.org/10.1007/s11097-010-9185-8

107. Menary, R., & Gillett, A. J. (2021). Are Markov Blankets real and does it matter? In D. Mendonça, M. Curado & S. S. Gouveia (eds.). *The Philosophy and Science of Predictive Processing*. Bloomsbury Academic (pp. 39-58).
108. Menzies, P., & List, C. (2010). The causal autonomy of the special sciences. In Macdonald, C., & Macdonald, G. (eds.), *Emergence in mind*, 108-128. Oxford: Oxford University Press.
109. Merker, B., Williford, K., & Rudrauf, D. (2022a). The integrated information theory of consciousness: a case of mistaken identity. Behavioral and Brain Sciences, 45, e41. http://doi.org/10.1017/S0140525X21000881
110. Merker, B., Williford, K., & Rudrauf, D. (2022b). The integrated information theory of consciousness: Unmasked and identified. Behavioral and Brain Sciences, 45, e65. http://doi.org/10.1017/S0140525X21002387
111. Merleau-Ponty, M. (1962). *The Phenomenology of Perception*. London: Routledge.
112. Millikan, R. G. (1984). *Language, thought, and other biological categories*. Cambridge, MA: MIT Press.
113. Millikan, R. G. (1989). In defense of proper functions. *Philosophy of science*, 56(2), 288-302. http://doi.org/10.1086/289488
114. Nguyen, J. (2016). On the pragmatic equivalence between representing data and phenomena. *Philosophy of Science*, 83(2), 171-191. https://doi.org/10.1086/684959
115. Nguyen, J., & Frigg, R. (2021). Mathematics is not the only language in the book of nature. *Synthese*, 198(24), 5941-5962. https://doi.org/10.1007/s11229-017-1526-5
116. Oppenheim, P., and Putnam, H. 1958. The Unity of Science as a Working Hypothesis. In Feigl, H., Scriven, M. & Maxwell, G. (eds.) *Minnesota Studies in the Philosophy of Science*. Minneapolis: University of Minnesota Press: 2: 3-37.
117. Oudiette, D., Leu, S., Pottier, M., Buzare, M. A., Brion, A., & Arnulf, I. (2009). Dreamlike mentations during sleepwalking and sleep terrors in adults. *Sleep*, 32(12), 1621-1627. https://doi.org/10.1093/sleep/32.12.1621
118. Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active inference: the free energy principle in mind, brain, and behavior*. Cambridge, MA: MIT Press.

119. Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

120. Pero, F., & Suárez, M. (2016). Varieties of misrepresentation and homomorphism. *European Journal for Philosophy of Science*, 6(1), 71-90. https://doi.org/10.1007/s13194-015-0125-x

121. Pigorini, A., Sarasso, S., Proserpio, P., Szymanski, C., Arnulfo, G., Casarotto, S., Fecchio, M., Rosanova, M., Mariotti, M., Lo Russo, G., Palva, J. M., Nobili, L., & Massimini, M. (2015). Bistability breaks-off deterministic responses to intracortical stimulation during non-REM sleep. *Neuroimage*, 112, 105-113. https://doi.org/10.1016/j.neuroimage.2015.02.056

122. Pillmann, F. (2009). Complex dream-enacting behaviour in sleepwalking. *Psychosomatic medicine*, 71(2), 231-234. https://doi.org/10.1097/psy.0b013e318190772e

123. Pincock, C. (2004). A new perspective on the problem of applying mathematics. *Philosophia Mathematica*, 12(3), 135–161. https://doi.org/10.1093/philmat/12.2.135

124. Pisano, N. A. (2023). An instrumentalist take on the models of the Free-Energy Principle. *Synthese*, 201(4), 126. https://doi.org/10.1007/s11229-023-04111-8

125. Price, H. 2011. *Naturalism without mirrors*. Oxford: Oxford University Press.

126. Raja, V., Valluri, D., Baggs, E., Chemero, A., & Anderson, M. L. (2021). The Markov blanket trick: On the scope of the free energy principle and active inference. *Physics of Life Reviews*, *39*, 49-72. https://doi.org/10.1016/j.plrev.2021.09.001

127. Ramstead, M. J., Kirchhoff, M. D., & Friston, K. J. (2020). A tale of two densities: Active inference is enactive inference. *Adaptive Behavior*, 28(4), 225-239. https://doi.org/10.1177%2F1059712319862774

128. Ross, C. A., & Ness, L. (2010). Symptom patterns in dissociative identity disorder patients and the general population. *Journal of Trauma & Dissociation*, *11*(4), 458-468. https://doi.org/10.1080/15299732.2010.495939

129. Rowlands, M. (2009). Extended cognition and the mark of the cognitive. *Philosophical Psychology*, 22(1), 1-19. https://doi.org/10.1080/09515080802703620

130. Rowlands, M. J. (2010). *The new science of the mind: From extended mind to embodied phenomenology*. Cambridge, MA: MIT Press.
131. Saigusa, T., Tero, A., Nakagaki, T., & Kuramoto, Y. (2008). Amoebae anticipate periodic events. *Physical review letters*, 100(1), 018101. https://doi.org/10.1103/PhysRevLett.100.018101
132. Schenck, C. H., & Mahowald, M. W. (1994). Review of nocturnal sleep-related eating disorders. *International Journal of Eating Disorders*, 15(4), 343-356. https://doi.org/10.1002/eat.2260150405
133. Schwitzgebel, E. (2019), Introspection. *The Stanford Encyclopedia of Philosophy* Zalta (ed.), https://plato.stanford.edu/archives/win2019/entries/introspection/
134. Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, 3(3), 417-424. https://doi.org/10.1017/S0140525X00005756
135. Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal* (27)379-423, 623-656.
136. Siclari, F., & Tononi, G. (2017). Local aspects of sleep and wakefulness. *Current Opinion in Neurobiology*, 44, 222-227. https://doi.org/10.1016/j.conb.2017.05.008
137. Sider, T. (2003). What's so bad about overdetermination?. *Philosophy and Phenomenological Research*, 67(3), 719-726. https://doi.org/10.1111/j.1933-1592.2003.tb00321.x
138. Smart, J. J. (1959). Sensations and brain processes. *The Philosophical Review*, 68(2), 141-156. https://doi.org/10.2307/2182164
139. Suárez, M. (2003). Scientific representation: Against similarity and isomorphism. *International studies in the philosophy of science*, 17(3), 225-244. https://doi.org/10.1080/0269859032000169442
140. Terzaghi, M., Sartori, I., Tassi, L., Didato, G., Rustioni, V., LoRusso, G., Manni, R., & Nobili, L. (2009). Evidence of dissociated arousal states during NREM parasomnia from an intracerebral neurophysiological study. *Sleep*, 32(3), 409-412. https://doi.org/10.1093/sleep/32.3.409
141. Terzaghi, M., Sartori, I., Tassi, L., Rustioni, V., Proserpio, P., Lorusso, G., Manni, R., & Nobili, L. (2012). Dissociated local arousal states underlying essential

clinical features of non-rapid eye movement arousal parasomnia: An intracerebral stereo-electroencephalographic study. *Journal of sleep research*, 21(5), 502-506. https://doi.org/10.1111/j.1365-2869.2012.01003.x

142. Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind.* Cambridge, MA: Harvard University Press.
143. Tononi, G. (2004a). An information integration theory of consciousness. *BMC neuroscience*, 5(1), 1-22. https://doi.org/10.1186/1471-2202-5-42
144. Tononi, G. (2004b). Consciousness and the brain: Theoretical aspects. In: Adelman, G., & Smith, B. (eds.), *Encyclopedia of Neuroscience 3rd edition.* Elsevier.
145. Tononi, G. (2005). Consciousness, information integration, and the brain. *Progress in brain research*, 150, 109-126. https://doi.org/10.1016/s0079-6123(05)50009-8
146. Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. *The Biological Bulletin*, 215(3), 216-242. https://doi.org/10.2307/25470707
147. Tononi, G., & Koch, C. (2008). The Neural Correlates of Consciousness: An Update. *Annals of the New York Academy of Sciences*, *1124*(1), 239-261. https://doi.org/10.1196/annals.1440.004
148. Tononi, G., & Koch, C. (2015). Consciousness: here, there and everywhere?. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1668), 20140167. https://doi.org/10.1098/rstb.2014.0167
149. Tononi, G., & Massimini, M. (2008). Why does consciousness fade in early sleep?. *Annals of the New York Academy of Sciences*, *1129*(1), 330-334. https://doi.org/10.1196/annals.1417.024
150. Tononi, G. (2012). The integrated information theory of consciousness: an updated account. *Archives italiennes de biologie*, *150*(2/3), 56-90. https://doi.org/10.4449/aib.v149i5.1388
151. Van Duijn, M., Keijzer, F., & Franken, D. (2006). Principles of minimal cognition: Casting cognition as sensorimotor coordination. *Adaptive Behavior*, 14(2), 157-170. https://doi.org/10.1177/105971230601400207

152. van Es, T., & Hipólito, I. (2020). *Free-Energy Principle, Computationalism and Realism: a Tragedy*.

153. Van Es, T. (2021). Living models or life modelled? On the use of models in the free energy principle. *Adaptive Behavior*, 29(3), 315-329. https://doi.org/10.1177%2F1059712320918678

154. Van Es, T., & Kirchhoff, M. D. (2021). Between pebbles and organisms: weaving autonomy into the Markov blanket. *Synthese*, *199*(3), 6623-6644. https://doi.org/10.1007/s11229-021-03084-w

155. Van Fraassen, B. C. (1980). *The scientific image*. Oxford: Oxford University Press.

156. Van Fraassen, B. C. (2008). *Scientific representation: Paradoxes of perspective*. Oxford: Oxford University Press.

157. Van Riel, R. (2013). Identity, asymmetry, and the relevance of meanings for models of reduction. *The British Journal for the Philosophy of Science* 64(4), 757-761. https://doi.org/10.1093/bjps/axs028

158. Weisberg, M. (2006). Forty years of 'The strategy': Levins on model building and idealization. *Biology and Philosophy*, 21(5), 623-645. https://doi.org/10.1007/s10539-006-9051-9

159. Weisberg, M. (2007). Three kinds of idealization. *The Journal of Philosophy*, 104(12), 639-659. https://doi.org/10.5840/jphil20071041240

160. Weisberg, M. (2013). *Simulation and Similarity: Using Models to Understand the World.* Oxford: Oxford University Press.

161. Wheeler, M. (2010). In defense of extended functionalism. In Menary, R. (ed.), *The Extended Mind*, 67-85. Cambridge, MA: MIT Press.

162. Wheeler M. (2019). Breaking the waves. In M. Colombo, E. Irvine, & M. Stapleton (eds.), *Andy Clark and His Critics*. Oxford University Press (pp. 81-95)

163. Wimsatt, W. C. (1994). The Ontology of Complex Systems: Levels of Organization, Perspectives, and Causal Thickets. *Canadian Journal of Philosophy Supplementary Volume*, 20: 207-274. https://doi.org/10.1080/00455091.1994.10717400

164. Woodward, J. (2003). *Making things happen: A theory of causal explanation.* Oxford: Oxford University Press.

165. Woodward, J. (2008). Mental causation and neural mechanisms. In Hohwy, J., & Kallestrup, J. (eds.), *Being reduced: New essays on reduction, explanation, and causation*, 218-262. Oxford: Oxford University Press.

166. Zadra, A., Desautels, A., Petit, D., & Montplaisir, J. (2013). Somnambulism: clinical aspects and pathophysiological hypotheses. *The Lancet Neurology*, 12(3), 285-294. https://doi.org/10.1016/s1474-4422(12)70322-8

167. Zadra, A., & Levitin, D. J. (2022). The disintegrated theory of consciousness: Sleep, waking, and meta-awareness. Behavioral and Brain Sciences, 45, e64. https://doi.org/10.1017/s0140525x21001850

168. Zahavi, D., & Parnas, J. (1998). Phenomenal consciousness and self-awareness: A phenomenological critique of representational theory. *Journal of Consciousness Studies*, 5(56), 687-705.

169. Zhong, L. (2014). Sophisticated exclusion and sophisticated causation. *The Journal of Philosophy*, 111(7), 341-360. https://doi.org/10.5840/jphil2014111724