Deflationary Realism:

Representation and Idealisation in Cognitive Science

Dimitri Coelho Mollo

Science of Intelligence Cluster & Berlin School of Mind and Brain

& Institut für Philosophie, Humboldt-Universität zu Berlin

Submitted manuscript – Please cite or quote only the published version

Accepted version forthcoming in *Mind & Language*

Abstract

Debate on the nature of representation in cognitive systems tends to oscillate between robustly realist views and various anti-realist options. I defend an alternative view, deflationary realism, which sees cognitive representation as an offshoot of the extended application to cognitive systems of an explanatory model whose primary domain is public representation use. This extended application, justified by a common explanatory target, embodies idealisations, partial mismatches between model and reality. By seeing representation as part of an idealised model, deflationary realism avoids the problems with robust realist views, whilst keeping allegiance to realism.

Keywords: Cognitive Representation; Naturalising Intentionality; Scientific Models; Deflationary Realism; Idealisation; Indeterminacy of Content

1 Introduction

A guiding principle that has marked most of cognitive science since its birth about 70 years ago is the idea that cognition is made possible, at least to a large extent, by internal computational processes that manipulate internal states that bear representational content. Much philosophical work has been dedicated to exploring the notions of representation and computation, with an eye especially to accounting for them in purely naturalistic, scientifically-acceptable ways. This computational-representational framework is no longer the 'only game in town', with interesting alternatives being developed by proponents of embodied and enacted cognition that often downsize, and sometimes eliminate, appeal to representation and/or computation in explaining cognitive capacities. At any rate, that framework plays to this day an important role in informing contemporary research in the cognitive sciences, and in light of this continuing relevance, philosophical interest in the nature of cognitive representation and computation goes on.

In this paper I focus on the notion of representation in cognitive science. Philosophical work on representation has blossomed in the '80s and '90s, but since the turn of the century interest seems to have waned (Godfrey-Smith 2006). The main candidates for a naturalistic theory of representation — informational semantics, inferential role semantics, and teleosemantics — all seem to face crucial difficulties; and the solutions put forward have failed to convince all, resulting in a lack of consensus and diminishing trust on the prospects of success of the naturalisation project. In the past years, however, hope has returned, and there has been a renewed interest in shedding light on the notion of representation. Some provide fresh developments of teleo-based theories (Neander 2017, Millikan 2017, Shea 2018); while others have aimed at exploring alternative views of the nature and role of representation in the cognitive sciences (Godfrey-Smith 2006, Sprevak 2013, Egan 2014, Coelho Mollo 2020).

Here I will follow the latter path, suggesting a different way of understanding cognit-

ive representations in light of recent philosophical work on the nature and role of models and idealisations in science. In a nutshell, I will argue that cognitive representations are part of an idealised explanatory model, the representational model, whose primary domain of application is the practices surrounding the use of external representations. Despite containing idealisations in its application to cognitive phenomena, I hold that the representational model captures actual features of cognitive systems: features that are relevant for explaining certain patterns of behaviour that call for representational explanation, which I call representational patterns. These considerations invite a realist view of representation, although one in which troublesome commitments of traditional theories are seen as idealisations, thereby deflating their metaphysical import. I call the resulting view 'deflationary realism'.

Here is how I will proceed in what follows. In section §2 I provide a rough taxonomy of the main views on the ontological status of representation. I explore the features of the representational model in section §3, arguing that representational explanation has its origins in the social practices sustaining public representation use, and that it has a privileged explanatory target, i.e. representational patterns. In section §4 I analyse the limits of the representational model when applied to cognitive systems. I put forward my alternative view of representation, deflationary realism, in section §5, and I argue that it avoids the problems with traditional views of representation. Finally, I contrast it with other recent deflationary approaches, content pragmatism and fictionalism, as well as with robust realist views, showing that deflationary realism is a distinct view (section §6).

2 Representation in Cognition

A promising approach to representations is to see them as theoretical posits that play an explanatory role in the cognitive sciences (Cummins 1989, Coelho Mollo 2015, Shea 2018). In the cognitive sciences, representations are typically seen as internal, often

subpersonal states that are about entities and processes in the world, in a way not too dissimilar to how pictures are about what is depicted, words are about what they refer to, and maps are about the landscapes they map. Representational contents place conditions on how the world is or should be, if the representation is, in a loose sense, to be appropriate. By being about things in the world, representations stand in for what they represent, and by operating over their own representational states, cognitive systems are able to produce behaviour that is appropriate to the circumstances the organism finds itself in. The physical realisations of representational states are representational vehicles: the physical states that bear representational content. In the case of the brain, representational vehicles are ultimately states and processes going on in neurons and/or in populations of neurons (and possibly in glia), although this does not mean that explanations in representational terms need be cashed out in neurophysiological terms [1].

A crucial contrast that sets apart different approaches to representation in the cognitive sciences is between views that insist that cognitive representations are theoretical posits that refer to real, objective entities in cognitive systems; and views that deny it. I will refer to the former as realist views, whilst I will call the latter anti-realist views.

Cognitive representations, by the realist's lights, really exist within cognitive systems. The notion is typically posited in the course of attempts to explain cognitive phenomena and intelligent behaviour, and its explanatory value is supposed to hinge on its corresponding to features that cognitive systems possess, and would possess even in the absence of any sentient being having cognitive states about those features, and independently of the explanatory purposes and practices in which those notions are employed. For anti-realists, in contrast, the notion essentially depends on conceptual apparatus that we as cognitive scientists or users of folk psychology impose on the world, and which is not committed to their being faithful to its nature and organisation.

[1] For recent debate and defence of the explanatory autonomy of representational, and cognitive explanation more generally, see the papers collected in Kaplan (2017).

Some of the most influential realist theories of representation belong to a family of views that I call robust realism. Robust realists aim at naturalising representation, explaining the notion in purely naturalistic and non-representational terms. They take some inner states of cognitive systems to have representational content due to their standing in some special natural relation(s) to what they represent, which bestows fairly determinate contents on them. Representations are taken to be structures in the cognitive system whose boundaries are relatively clearly definable, stable, repeatable (i.e. participate in different cognitive processes whilst preserving the same content), and often composable (i.e. that can be systematically combined with other representations to yield more complex representations). Robust realists are also committed to there being identifiable subsystems whose role it is to use cognitive representations: representation consumers. Informational and teleoinformational semantics (Dretske 1981, Neander 2017), teleosemantics (Millikan 2017), inferential role semantics (Block 1986), and structural representation (Ramsey 2007, Shea 2018) are examples of robust realist theories.

Robust realist theories of representation face several objections, and despite decades of work dedicated to tackling them, it is still unclear whether they can be satisfactorily addressed. Perhaps the most damning are the various objections from indeterminacy of content. Briefly, the worry goes that robust realist theories yield representations with wildly indeterminate, non-unique contents. Representations would thereby have several simultaneous contents, many of which unrelated to or of a different kind than the content that would seem as the most intuitive and/or explanatorily apt. Indeterminacy of content is an unfortunate result for several reasons. First, it seems strongly counterintuitive when applied to conscious representational states, as they strike us as having determinate contents[2]. More importantly, indeterminacy of content jeopardises the explanatory

[2]Determinacy of content is not the opposite of vagueness. Determinate contents may include vagueness, for instance if they involve vague concepts, or if the kinds they are about are fuzzy (Millikan 2017).

role representations are supposed to play, risking to make representational explanation vacuous or uninformative, and misrepresentation hard to accommodate, or completely impossible[3]. A less discussed difficulty for robust realist theories is their commitment to identifiable and relatively stable representational vehicles and representation consumers. Issues with vehicle and consumer individuation surface especially in teleologically based approaches, which currently appear to be the most promising family of robust realist views (Godfrey-Smith 2006, Cao 2012)[4].

On the anti-realist side, a variety of positions withhold ontological commitment to representations on often different grounds. Some believe that representations do not exist, or if they do, that they play only a minor explanatory role in the cognitive sciences, being fruitfully replaced by alternative theoretical posits coming from dynamical systems theory and neurophysiology (Van Gelder 1995), and/or from alternative, embodied approaches to cognition (Thompson 2007, Hutto & Myin 2013). I remain neutral on the prospects of success of anti-representationalist views. My aim in the foregoing is to try and provide a more satisfactory version of representationalism: deflationary realism. My view will then have to earn its keep, theoretically and empirically, in comparison to anti-representationalist alternatives — although, as I point out later (section §7), deflationary realism removes some of the grounds for seeing the two camps as stark competitors.

Other forms of anti-realism about representation reject ontological commitment to representations, but hold that nonetheless the notion plays central and/or ineliminable roles in the cognitive sciences. My deflationary realism bears interesting relationships to this family of positions, whose most prominent candidates are content pragmatism, and representational fictionalism.

Content pragmatism sees cognitive representation as part of a gloss that has heur-

[3]Three indeterminacy problems have been most discussed: the disjunction problem (Fodor 1987), the distality problem (Dretske 1986), and the functional indeterminacy problem (Fodor 1990). See Ryder (2009) for an overview, and Neander (2017, chap. 7) for other types of indeterminacy problems.

[4]This is particularly problematic for consumer-based approaches to teleosemantics, which see representation consumers as helping determine content. In these accounts, difficulties individuating consumers infect content determination, being an additional source of indeterminacy.

istic value, but that makes no ontological commitment to representations (Egan 2014, Coelho Mollo 2020). On this account, cognitive representations are not posits of cognitive theories proper, but are rather part of the extra-theoretical apparatus that allows us to connect purely computational explanations of cognition to the externally-individuated tasks we are interested in explaining. On the other hand, according to representational fictionalism talk of cognitive representation is a useful fiction (Sprevak 2013), or a form of pretence (Toon 2016): when making claims about cognitive representations, scientists engage in fictional discourse — a discourse that works as a useful device for understanding behaviour, but that does not commit us to the reality of cognitive representations. A central difficulty for both pragmatism and fictionalism is to justify the claim that representational talk is not ontologically-committing, despite being part and parcel of our best explanations in cognitive science — in other words, they seem to flout the principle of inference to the best explanation (Sprevak 2013). I believe that deflationary realism accommodates some of the insights that motivate pragmatism and fictionalism, while keeping at bay their problems by sticking to a less-than-robust form of realism. I compare the three views in section §6.

3 Representational Practices, Representational Patterns

A particularly fruitful way of understanding the explanatory appeal of the notion of representation in cognitive science is by recognising its close ties to a model of explanation that has its original application in the social practices sustaining the use of public representations (Godfrey-Smith 2006). This 'representational model' is originally targeted at capturing some of the central features that characterise the use of public symbols by agents in social interactions, and in everyday life more generally. Clear examples include linguistic symbols, traffic signs, and city maps.

Representations, on this model, are public objects that agents use in guiding thinking and behaviour toward something else. Representations act as stand-ins thanks to some

property they have that makes them into adequate guides, for specific purposes, and by agents with the adequate interpretative capacities, to thinking and behaving toward what is represented. Some typical features of public representations characterise the representational model: representations are distinct from the agents that use them, agents intentionally use them as stand-ins for specific entities in specific circumstances, and there is some relation between the representation and what it is used to represent that, when correctly interpreted, endows them with determinate contents, allowing this usage to work adequately enough.

Take a map of Berlin. If it is accurate enough, I can use it to navigate Berlin successfully, even if I had never been to the city before. The map allows me to move around Berlin in line with my varying interests and needs during the trip. I may check the map so as to get to the Pergamon Museum, which stands in my list of attractions to visit. The map allows me to reach the museum from a variety of different starting points: a handy property for travellers like me, who enjoy wandering aimlessly around a new city. It also allows me to reach the museum regardless of variations in other, non-spatial aspects of my context, such as meteorological conditions, means of transportation (if it is raining, I may prefer taking public transport rather than walking), and even my desires and needs (I may want to go to the museum not for the artworks, but rather because I am meeting a friend there, or because I have been told that the café is excellent and I am hungry, etc.).

It so happens that I am actually quite familiar with Berlin, and I can reach the Pergamon Museum successfully and reliably across many different contexts without the need of a map, or a list of directions, or any other kind of external representation. A natural move then is to explain this capacity of mine in terms of an internal representation, perhaps even one that is to some extent similar to a map — e.g. a 'cognitive map' in entorhinal cortex (Moser et al. 2008) — that cognitive subsystems use to guide my behaviour. This is an illustration of the representational model being applied to explain

a cognitive capacity. The approach consists, roughly, in seeing (some) cognitive states as guides to states and events in the world, like external representations — with intentional agents replaced by non-intentional cognitive processes and subsystems that use them in the appropriate ways (Godfrey-Smith 2006). Application of the representational model to understanding cognition leads to attempts to individuate the corresponding parts of the model in cognitive systems. We try, that is, to find in cognitive systems states and processes that play roles analogous to those of external representations, and of the agents that interpret and use them. It becomes thus vital to explain how the representational model can work without making reference to intentional agents and social interpretative practices. This is the core challenge that realist, naturalistic theories of representation try to meet: explaining by non-intentional and non-semantic means how cognitive states acquire content, and in such a way that they can be used as stand-ins by other cognitive processes (for which appeal to social practices of interpretation is out-of-bounds).

Importantly, the distinctive explanatory target of representational explanation, and its explanatory role in the cognitive sciences more generally, constrains the ways in which the representational model is applied to cognitive systems. Representation comes to the fore when the targets of explanation are certain types of patterns of behaviour, which we tend to see as intelligent and for which appeal to representation is distinctively explanatory — i.e. for which appeal to representation leads to better explanations than would be possible otherwise (Shea 2018, Rupert 2018). I call these explanatory targets 'representational patterns'. In order to individuate more precisely the domain of application of the representational model in the cognitive sciences, we have to look at the features that its explanatory targets possess in its primary domain of application, that is to say, the features of the distinctive behavioural patterns that the social practices tied to the use of public representations make possible.

As we have seen, public representations are objects used as guides for thinking or behaving toward something else. A city map allows its user to navigate a city, reach-

ing desired destinations from a variety of different starting points, and regardless of variations in non-spatial aspects of the situation. The map thereby enables reliable behaviour, as it allows one to reach desired locations, and it does so robustly and flexibly, inasmuch as one can reach them from different points in the city, for different reasons, and under different environmental conditions. Use of the map leads to reliable and robust behavioural outcomes[5]. This, I believe, is one of the characteristic features of public representations. They enable reliable, robust and flexible interaction between agents and aspects of the world across a variety of circumstances. They do so in complicated ways, as the complexity of cartographic representations — which contain approximate spatial mappings, icons, conventional symbols, etc. — illustrates (Camp 2007, Rescorla 2009).

If one knows a city well, one can navigate the city reliably across many different contexts without the need of a map, or a list of directions, or any other kind of external representation. The *explanandum* in this case, i.e. a certain pattern of (navigational) behaviour, is sufficiently analogous to the one in the external map case to warrant the application of the representational model, positing internal states and processes that function analogously to external maps, i.e. cognitive maps[6].

Representations help explain patterns of robust, reliable, flexible behaviour due to their standing in for what they represent, allowing 'surrogative reasoning' (Swoyer 1991). Representations also help explain why in some cases behaviour reliably fails by appeal to the falsity or inaccuracy of the representations used. If a map of Berlin wrongly shows the Pergamon Museum to be in the Tiergarten park, rather than on the Museum

[5]The degree and kind of reliability, success, and robustness varies depending on the type of representation and its level of accuracy. A list of written directions from point A to point B is reliable only from one point of departure, so it is not much robust (although it is robust across non-spatial contextual variation).

[6]It may be argued that the similarity between the two cases stems from both involving the use of a cognitive map, in one case prompted or assisted by an external map, and in the other without such an external prop. This worry is unwarranted, as the two cases are importantly different. The internal representations built while using an external map are partial and parasitic on the use of the external prop, and behaviourally useless without it — an instance of cognitive offloading. In contrast, when navigating a well-known city without the aid of an external map, successful behaviour is enabled by rich internal representations that play an analogous guiding role.

Island, it will reliably and robustly lead its users to fail to reach the museum if that is their desired destination (if they don't know better, they will rather robustly and reliably reach the Tiergarten park) — analogously if the inaccurate map is some sort of internal representation 'consulted' by navigation systems. Correct representation helps explain successful behaviour — that enters in a representational pattern — and incorrect representation helps explain unsuccessful behaviour — that patterns in the characteristically robust and reliable way. This explanatory role is common to both public and cognitive representations.

Representational patterns consist in the behavioural regularities underscored by robust, flexible and reliable interactions between complex organisms and their ever-changing environments across time. Representational explanation is particularly adequate to capturing such behavioural regularities, even in cases in which explanation in terms of use of external representations is not possible. These distinctive patterns of behaviour invite the application of the representational model in so far as they bear a striking similarity to the feats that we, as intentional agents, can accomplish by means of using public representations. Internal representations can then be appealed to in explaining those behavioural patterns, in analogy to public representations and the behavioural patterns they make possible[7].

These considerations make clear that, while the label 'representational pattern' may suggest that this characterisation is circular or uninformative, this is not the case. The distinctive properties of representational patterns — i.e. behavioural robustness, flexibility and reliability — do not presuppose representations, but rather suggest representations as the suitable theoretical posits able to explain how those properties come about. In other words, representational patterns are types of pattern of behaviour that, when taken as *explananda*, call for positing representations as their *explanantia*. (Compare:

[7]This is not to say that representational explanation cannot be applied to other kinds of behavioural pattern, such as context-insensitive stereotypical behaviour. The claim is rather that when applied to behaviours other than those that make up representational patterns, representational explanation may be less adequate, and it may well play only a heuristic or pragmatic role.

many diseases exhibit patterns of symptoms and patterns of transmission that are best explained by appeal to the action of microbes such as bacteria and viruses, but surely the germ theory of disease was neither circular nor uninformative even before microbes could be observed.)

More schematically, the main point of this section can be roughly summarised thus (see also Godfrey-Smith 2006):

1. Agents display patterns of behaviour marked by a high degree of robustness, reliability, and flexibility.

2. Some of these patterns of behaviour are best explained by the use of external representations by intentional agents; such explanations have specific features that can be seen as composing a (representational) model of explanation.

3. Other similar patterns of behaviour cannot be explained by the use of external representations.

4. Given their similarity to the patterns of behaviour explained by external representations, it is plausible that the adequate model for explaining them is also the representational model.

5. Therefore, we have good reason to apply the representational model to cognitive systems by positing internal representations whose nature and function is analogous to that of external representations.

4 Limits of the Representational Model

As I argued in the previous section, once our aim is to explain certain kinds of patterns of behaviour — which I have been calling representational patterns — application of the representational model of explanation seems to be a promising way to go. Applying the representational model to help explain those patterns where public representations are

not in question, as in many cases of human and non-human cognition, suggests positing states and processes internal to cognitive systems that play roles analogous to those of public representations and their users: cognitive representations and representation consumers. Even though recent work in embodied cognition offers alternative, often non-representational models for at least some cognitive capacities, application of the representational model to cognition has proved to be scientifically fruitful, spawning several fields of research that have been making progress to this day (see Kriegeskorte & Douglas 2018 for a recent review).

There are however features of public representations that are difficult to square with how states and processes in cognitive systems work. These are the places in which the representational model fails to be faithful to its extended domain of application; where its limitations when applied to the cognitive sciences come to the fore. These are also the places that invite caution before turning elements of the model into requirements on the metaphysics of cognitive representation. I do not purport here to give anything close to a theory of public representations, or of the social practices that underlie their use. For our purposes it suffices to identify some of their central features so as to see whether and to what extent they can be shared by cognitive representations. I hope that the features I will mention are relatively uncontroversial, requiring the adoption of no specific theory of public representational practices.

First, public representations depend at least partly on intentional agents that take them to be representations, use them as representations, and are able to interpret them appropriately. Public representations are also typically intended to function as representations by their creators. If one wants to stay within a naturalistic approach, appeal to agents' intentional states in explaining cognitive representation is off the cards[8]. Second, public representations rely on the existence of social practices of interpretation,

[8]This is so even in naturalistic social- or language-based views, which see intentionality as belonging primarily to external symbols, and only derivatively to internal states. These accounts typically base the intentionality of external symbols on non-intentional mechanisms of social conformity (Haugeland 1990, Cash 2009).

which allow members of a culture to extract the intended meaning from representations. Since I do not read Japanese, I cannot adequately interpret Japanese linguistic symbols. There are also less obvious cases: 'reading' some kinds of maps, graphs, mathematical notations, not to mention pieces of art, requires interpretive skills that rely on specific practices maintained and transmitted by communities. This is another aspect that cognitive representations cannot share with their public counterparts[9].

The traditional solution to these disanalogies is to transform the agents and interpretive practices that underlie public representation use into very simple input-output cognitive mechanisms following (typically) computational rules (Dennett 1978, Godfrey-Smith 2006). Such simple mechanisms 'use and interpret' only in an extremely watered down sense: there is no intentionality involved, just the automatic workings of (computational) mechanisms. While this 'homuncular' strategy seems to accommodate the agential and interpretive elements of the representational model in its application to cognitive systems, there are reasons to be less confident about the details.

Public representations typically have relatively clear boundaries in space and/or time, at least given appropriate interpretive practices. A sentence has a clear beginning and end (and clear parts), a city map can fit in one's pocket, and pictures can be framed and affixed to the wall. Public representations are typically spatially and temporally delimitable, often portable, and clearly separable from the agent that makes use of them. In the case of biological cognitive systems, it is doubtful that there are any neat boundaries to representational vehicles. Candidate vehicles are features of neural activity in cortical regions, whose representational 'code' is still unclear (deCharms & Zador 2000), as are their boundaries: where the representation begins and ends is a question with a much less straightforward answer in the case of cognitive systems, given the complexity of their connectivity structure, and the ubiquity of feedback loops (Cao

[9]Unless one opts for a social- or language-based view of intentionality. At any rate, under such a view the practices that underlie representational content must be non-intentional, so as to keep allegiance to naturalism.

2012, Artiga 2016). This is not to say that it cannot be answered; and in some sensorimotor areas there are populations of neurons that allow some degree of delimitation as representational vehicles, such as cortical columns in V1.

However, even in their case the situation is rather complicated (Cao 2019). Such populations have preferred stimuli to which they respond, but they also respond to other stimuli, although less often and less strongly. Moreover, inhibitory connections to other populations of neurons play a role in their workings. Should we include populations of neurons that respond to a certain type of stimulus non-preferentially as part of the representation? Or what about the populations involved in inhibitory activity? Or the populations that block or allow such inhibitory activity to go through? And this without even mentioning neural plasticity, degeneracy and pluripotentiality, and network reorganisation, which add further layers of complexity and instability to vehicle individuation (Anderson 2014). In other words, individuating representations and their parts in cognitive systems is rather difficult, and there is no reason to believe that cognitive representational vehicles can be bounded, even approximately, as public representations can.

Relatedly, the clear separation between representation and user that characterises public representations is hardly applicable to cognitive systems. Although the assumption of separability plays a central role especially in teleosemantic theories, where the distinction between representation producer and consumer helps make the approach get off the ground, it is unclear whether there is any support for such distinction in cognitive systems (Godfrey-Smith 2006, Cao 2012). For reasons analogous to the above, not only are the boundaries of representational vehicles far from clear, but so are also the boundaries of the putative subsystems that use them as representations.

As we have seen, part of what helps fix the content of public representations are the interpretive practices socially produced and transmitted in a community. These practices enable users correctly to interpret the content of public representations, using

them in ways appropriate to what they represent, even in cases where the relationship between representation and what is represented is very convoluted and opaque, as in natural language. Public representations typically have determinate, or fairly determinate content because interpretive practices allow representation users appropriately to grasp the intended content of the representation.

In the case of cognitive representations social interpretive (intentional) practices cannot be appealed to in helping fix their content, if we are to keep allegiance to naturalism. It is to this problem that robust realist theories of representation have dedicated most of their attention. They propose natural substitutes to the interpretive practices that sustain public representation, in an attempt to show that natural relations and processes can play a similar role in helping fix the content of cognitive representations. Most of the debate has focused on which combinations of natural relations and processes are able to bestow fairly determinate content on representations, as interpretive practices do in the case of public representations. The failures of existing proposals in this regard are illustrated by the many indeterminacy problems that have surfaced in the literature.

In brief, application of the representational model to cognitive systems has important limitations, if taken literally. There are at least three central posits of the representational model that do not match known properties of cognitive systems or that are difficult to account for: clearly bounded representational vehicles, the separability of representation and consumer, and determinacy of content. These mismatches underlie some of the crucial objections moved against robust realist theories of representation. Robust realist theories take the representational model seriously, and claim that it is at least in principle, if not in (future) practice, possible to map its posits into actual entities and processes in cognitive systems. The failure of such mapping, however, need not leads us to anti-realism about representation. In the rest of this paper, I explore a view of representation at odds with both robust realism and anti-realism: deflationary realism.

5 Deflationary Realism

Deflationary realism contends that the representational model, when applied to cognition, is an idealised scientific model. In so far as it embodies idealisations, the model incorporates falsities. But it does so in order to bring forth the central causal factors that make the phenomena I have been calling representational patterns possible. Cognitive representation, as a theoretical posit of the model, approximates real properties of cognitive systems, despite the idealisations it involves.

Contemporary philosophy of science increasingly recognises the significance of models and idealisations in science, and there is a rich ongoing debate about their nature and role (Weisberg 2013, Morrison 2015, Potochnik 2017). My understanding of scientific idealisation is close to Potochnik's (2017): idealisations are distorted representations of aspects of a target system, which by means of their partial falsity bring forward focal causal features of the system relevant to explaining a phenomenon of interest. For instance, an evolutionary model may assume infinite population size (an idealisation) to bring into focus the specific contributions made by natural selection to the evolution of traits within a finite population, disregarding in this way other causal influences such as genetic drift. Idealisations are, for Potochnik, inevitable in science, given the complexity of the world, our explanatory aims, and the limitations of our cognitive capacities and epistemic reach.

Idealised models permeate science: they are essential for achieving central scientific epistemic aims, such as explanation, prediction, and understanding (Bokulich 2011, Potochnik 2017). Idealised models are explanatory (at least) in so far as they truly represent, thanks to the simplifications and distortions they embody, focal patterns of (causal) dependency that are relevant for explaining a phenomenon of interest. Idealisations are partly false, so that they can be partly true: the falsities they embody are instrumental for bringing forth actual causal structures of particular explanatory interest, which would be indiscernible otherwise, lost in the enormously intricate web of

causal dependencies and factors underlying most real phenomena.

Instead of a true/false dichotomy, which would force idealised models to be simply seen as false — and thus at most as useful fictions — we can more fruitfully consider the relationship of model to target partly in terms of degrees of similarity, with some models saying more true things than others about the modelled systems (Weisberg 2013, Morrison 2015). Alternatively, we can take idealised models to be true of the causal patterns they capture, while being false of phenomena as a whole, as they contain idealisations (Bokulich 2011, Potochnik 2017).

Deflationary realism holds that the cognitive representational model is an idealised model: a partially distorted and simplified picture of the causal features that contribute to bringing about the characteristic patterns of behaviour that it aims to explain, namely representational patterns. There are at least three grounds for taking the cognitive representational model to embody idealisations. First, as we have seen, it is a model that has its origins and primary motivation elsewhere. Application of the representational model to the cognitive sciences features limitations and shortcomings due to the fact that it is an extension of the application of concepts that belong to one domain — representational social practices — to a rather different one, that of cognitive states and processes. Second, cognitive systems are highly complex, and an indeterminate number of interacting causal factors are responsible for bringing about cognition and intelligent behaviour. Complex systems can be modelled only partially, and idealisation becomes crucial to make models tractable, understandable, and epistemically useful by simplifying some factors and ignoring others. Third, and relatedly, several of the relevant causal factors for explaining behaviour — e.g. fine-grained neural workings, evolutionary history — are either currently or in principle epistemically inaccessible to us. In consequence, we can do little else than to idealise them in our models, in light of our best guesses, or else idealise or abstract them away.

According to deflationary realism, application of the representational model to the

cognitive sciences need not, and should not, entail literal commitment to all of the entities and properties posited by the model. In so far as it embodies idealisations, a degree of mismatch between the representational model's commitments and the actual properties of cognitive systems is inevitable. I take this mismatch to underlie the philosophical *conundra* that have plagued robustly realist theories of representation. These arise not because the theories are incoherent or false, but rather because they fail to recognise the idealisations they embody. When these *conundra* are revealed as what they are, namely unsurprising offshoots of an idealised model, their metaphysical *gravitas* gives way to a healthier, epistemic modesty.

Seeing representational vehicles as bounded is a simplifying assumption, one that idealises away many of the intricate causal influences that parts of the cognitive system effect on each other. This allows us to bring into focus in a tractable way the crucial causal contributions that specific parts of the system make to the behaviour of interest. Similarly, the separability between representations and their consumers is an idealisation: it ignores the close interdependencies between most parts of the system, simplifying its organisation so as to bring to the fore a useful distinction between different functional roles; distinction that does not exist in such a neat way in the actual system, but that makes it more comprehensible and tractable. This also applies to types of explanation that focus on coarse-grained functional and computational levels of description of cognitive systems, which beside idealisation also make use of abstraction to bring into focus the relevant *explanantia*. These types of explanation tend to remain non-committal about the vehicles that realise the posited functional or computational structure, abstracting away from fine-grained physical details insofar as they are taken to be of little or no relevance to their explanatory aims. In short, they abstract away the details that impede appreciation of the explanatory causal pattern for the phenomena they seek to explain, positing neatly distinguishable and functionally/computationally bounded units and processes (Weiskopf 2017).

These aspects of the representational model work as idealisations, rather than as literally true of what is modelled. Even if it should turn out that such aspects of the representational model never strictly apply to cognitive systems, that is, that representational vehicles never have clear boundaries, and can never be fully separated from the subsystems that use them; they would still capture explanatorily relevant causal features of cognitive systems. Idealised (and abstract) models can be explanatory, as they reveal and bring into focus the subset of causal factors most relevant for explaining a phenomenon of interest.

These two idealisations embodied in the cognitive representational model are mostly due to the causal complexity of cognitive systems. Other idealisations may also be partly motivated by our epistemic limitations. This is the case, I take, of determinacy of content.

Deflationary realism invites a different take on the indeterminacy problems that plague robust realist theories. The naturalistic substitutes to the content-fixing interpretation practices of public representations are unlikely to be fully satisfactory replacements. The complexity of the factors that underlie representational patterns is plausibly such that no theory can fully account for them[10]. Many of the potentially relevant factors are moreover epistemically opaque to us — we have very little epistemic access to the details of the evolutionary history of cognitive systems, and we have little idea of what entities in cognitive systems are candidates for standing in the relevant naturalistic relations to things in the world. Given our epistemic limitations, our theories can only capture those factors in a rather coarse-grained way, giving us at most an idealised, partial grasp of the full story, which is though very often good enough for explanation and prediction.

According to deflationary realism, indeterminacy of content stems from the attempt to fit a complex system into a relatively simple model. The partial grasp on the relevant

[10]This is plausibly true of public representations as well. Much that calls for explanation is hidden in the superficially straightforward notion of 'interpretive practice'.

cognitive workings that existing theories of representation provide, their limited focus on one or some aspects that are relevant in bringing about representational patterns result in content indeterminacies. Indeterminacy of content is the result of the limitations of our models and of our knowledge; limitations that are unlikely to ever go away, given the complexity of the systems modelled, the epistemic opacity of the relevant factors, and the consequent partiality of our models. That indeterminacy of content marks all our attempts at giving a realist theory of representation is not the expression of a metaphysical problem, but rather a consequence of the adoption of an idealised model to explain certain phenomena. Incompleteness and opacity notwithstanding, such partial grasp over the workings of cognitive systems furnish helpful approximations of the core aspects of what makes representational patterns possible.

In many cases, furthermore, determinacy of content may be one of the elements of the representational model that match only very poorly its extended domain of application. Imposing the requirement that contents be fairly determinate on relatively simple systems, such as the toy example of frogs' tongue-snatching mechanism, can be misleading. There may be no grounds for insisting that the mechanism represents either 'fly', 'frog food', or 'moving black dot' — and similarly for other philosophical toy cases. These are cases in which theorists have tended to commit to elements of the representational model that do not approximate much anything in the systems under analysis.

The fact that the representational model, when applied to cognitive science, is an idealised model does not hurt its explanatory value in explaining representational patterns of behaviour. Ascribing determinate representational contents to internal states of cognitive systems helps explain certain patterns of behaviour, even if those contents are only idealisations. And even if the notion of representation only approximately captures the properties of the cognitive states and processes responsible for those patterns, it identifies the relevant causal dependencies for explaining the latter. Representational posits pick out some subset of the actual causal dependencies between entities and processes in

cognitive systems and their environments — those focal to explaining representational patterns — although without providing anything close to a true overall depiction. The cognitive representational model possesses a feature typical of explanatory models: it answers 'what-if-things-were-different' questions (Bokulich 2011, p. 43). Had the content of a representation as fixed by the model been different, the model would have predicted specific changes in downstream processing and behaviour in light of that difference. If these predictions are investigated and shown to be empirically adequate, they lend force to the claim that the model captures real causal dependencies relevant to cognitive functioning, despite, or rather thanks to, the idealisations it embodies.

In sum, the representational model largely works in explanations in the cognitive sciences because the model approximates many important features of actual cognitive systems, without though being a perfect fit. We should thereby be realists about cognitive representation, but in a way that concedes that what we are being realists about is only approximately what our model posits. This is a modest form of realism: one that recognises that we are on to something, and even have a pretty good idea of what that something looks like, whilst also recognising that our model has limitations, and that many aspects of the model cannot be precisely applied to the systems under investigation.

6 Deflationary Realism as a Distinct Position

Where does this picture leave us for what regards the ontological nature of representation? At first glance, deflationary realism may look like a form of pragmatism or fictionalism. We apply the representational model to cognitive systems because it is a useful tool to explain certain interesting phenomena. But this does not require or suggest that we should thereby commit to the actual existence of cognitive representations. They may be only fictions in an model, or part of a gloss which, albeit inevitable, does not justify ontological commitment to its posits. What makes this deflationary

model-based approach any different from those anti-realist positions?

Content pragmatism sees appeal to representation as part of an intentional gloss that, while helpful in so far as it allows theorists to track and grasp the relevance of internal computational states and process to the explanation of externalistically individuated tasks, is not part of the theory proper. The latter has recourse only to computational states and processes, representations being ascribed purely for heuristic reasons. In contrast, deflationary realism follows much contemporary philosophy of science in seeing models and idealisations as part and parcel of scientific theorising: science is permeated with idealisation, which is in most cases essential to theories, explanations, and predictions. As parts of an idealised model explanatory of certain specific patterns of behaviour, representations are part of cognitive theories proper, and not a gloss external to them.

The version of fictionalism closer to deflationary realism is Sprevak's (2013). In the picture he presents, representations are useful fictions, and scientific statements about representation are not aimed at truth, but rather, at best, at truth-in-the-fiction (i.e. in the idealised model). He even briefly considers understanding the pertinent kind of fiction as scientific idealisation (p. 17). Inasmuch as idealised models are not literally true of what they model, the representational model, when applied to cognitive systems, may be seen as a fiction (Frigg 2010). It may therefore be argued that deflationary realism is not really a form of realism, since what it invites ontological commitment to are not cognitive representations *per se* — as they are part of an idealised model — but rather different kinds of entity with which the latter share some properties. Cognitive representations would thus be useful fictions: fictions because they do not really exist, given their status as idealisations, and useful because they approximate actual states and processes well enough to allow prediction and intervention.

Deflationary realism resists this conclusion. It admits that some features of the representational model are idealisations, distortions of reality. But a proper understanding

of the nature of scientific idealisation suggests that this does not preclude the representational model from capturing actual causal features of cognitive systems. As we have seen, there are at least two ways of seeing the relations between idealised models and reality that warrant ontological commitment to their posits: we can either take idealised models to approximate truth well enough to justify ontological commitment to its posits (Weisberg 2013); or we can take idealised models to capture the actual causal features of the system that are most essential to explaining the phenomenon of interest (Bokulich 2011, Potochnik 2017). Accepting any of these views neatly distinguishes deflationary realism from anti-realist approaches: both tie idealisation to truth sufficiently strongly to resist the anti-realism about representations that fictionalism and pragmatism recommend. Idealised models can be explanatory and reveal true causal features of modelled systems.

Deflationary realism is superior to pragmatism and fictionalism inasmuch as it avoids their main shortcoming. By keeping allegiance to realism, deflationary realism makes clear why appeal to representation is so important, and possibly ineliminable, from the cognitive sciences — a feat that pragmatism and fictionalism struggle with. The representational model captures those real causal features of cognitive systems that are more directly causally relevant to explaining certain phenomena in the world: representational patterns. The causal role of cognitive representations remains thereby unscathed. Consequently, there is no challenge to scientific explanatory principles: the close tie between explanation and ontological commitment that the principle of inference to the best explanation dictates is not put in jeopardy.

Finally, although I have underlined the differences between deflationary realism and more robust realist accounts, I do not take this to indicate full incompatibility. Perhaps robust realist theories can be read as providing partial representational models, and the divergences between them as differences about which aspects of the causal structure of cognitive systems, and their history, are given more emphasis. Under this deflationary

light, the preferred natural relations that different robust realist theories appeal to are some of the relevant factors underscoring the existence of representational patterns, and consequently the value of representational explanation. This is the reading that deflationary realism recommends[11].

It is unlikely that this relatively modest view is what many robust realists are pursuing. The centrality given to indeterminacy of content problems, and the high stakes ascribed to these debates as saving or sinking the prospects of naturalising representation and thereby of a naturalistic cognitive science, suggest that the aims of these theories are more metaphysically ambitious. Moreover, debates between robust realists have mostly taken the form of a fight between competitors, while deflationary realism sees these theories as mutually compatible. To some extent deflationary realism is little more than an invitation for metaphysical modesty, and for giving a larger role to epistemic considerations when analysing the nature and role of cognitive representations. Modest versions of existing realist views along the lines above are therefore not at odds with deflationary realism, but are rather instances of it.

7 Concluding remarks

Before concluding, I would like to come back briefly to the dispute between representationalist and anti-representationalist views of cognition adumbrated in section §2. The view I have been proposing has it that the representational model is an idealised, partial model particularly suited to explaining certain patterns of behaviour. Given the

[11] This is in the spirit of Dennett's (1991) professedly mild realist view. However, it differs from it in at least three central aspects. First, Dennett's target are propositional attitudes, i.e. personal level states such as beliefs and desires, while the deflationary realism I propose targets mainly subpersonal states and processes relevant to the cognitive sciences. Second, Dennett relies on claims about the nature and features of what he calls real patterns, which need not be causal. In contrast, deflationary realism relies on the epistemic practices of modelling and idealising in science, which are aimed at revealing causal explanatory patterns and structures. Finally, he seems to suggest scepticism about internal representational and computational vehicles. While deflationary realism remains non-committal about many of the fine-grained features possessed by the states and processes that representations, as parts of idealised models, capture or approximate, it does not embrace scepticism about representational and computational vehicles.

complexity that marks cognitive systems and their interactions with their material and social environments, it is inevitable that different research aims and methods lead to a diversity of partial models, each bringing into focus one or a few explanatorily relevant factors, contributing a part of the story of how cognition and intelligent behaviour are brought about. Embodied and enactive approaches to cognition offer distinct models, in light of often different explanatory aims. Representationalism, once seen under the light that I recommend, is not incompatible with such approaches, which tend to downsize or eliminate appeal to representations. Rather, we have a rich variety of partial models, embodying different idealising assumptions and theoretical abstractions, all of which can be legitimately explanatory. Representationalism and anti-representationalism are thus not competitors, but candidates for pluralistic integration (Mitchell 2003)[12].

Acknowledgements

I am indebted to Nicholas Shea, Michael Pauen, Margherita Arcangeli and Matteo Colombo for helpful comments on previous versions of this paper.

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2002/1 "Science of Intelligence" – project number 390523135

References

Anderson, M. L. (2014), *After Phrenology: Neural Reuse and the Interactive Brain*, MIT Press.

Artiga, M. (2016), 'Teleosemantic modeling of cognitive representations', *Biology and Philosophy* **31**, 483–505.

Block, N. (1986), 'Advertisement for a semantics for psychology', *Midwest Studies in Philosophy* **X**(1), 615–78.

Bokulich, A. (2011), 'How scientific models can explain', *Synthese* **180**, 33–45.

Camp, E. (2007), 'Thinking with maps', *Philosophical Perspectives* **21**, 145–182.

[12]Though see Chirimuuta (forthcoming) for a less optimistic point of view.

Cao, R. (2012), 'A teleosemantic approach to information in the brain', *Biology and Philosophy* **27**, 49–71.

Cao, R. (2019), Computational explanations and neural coding, *in* M. Sprevak & M. Colombo, eds, 'The Routledge Handbook of the Computational Mind', Routledge, pp. 283–296.

Cash, M. (2009), 'Normativity is the mother of intention: Wittgenstein, normative practices and neurological representations', *New Ideas in Psychology* **27**, 133–47.

Chirimuuta, M. (forthcoming), Charting the heraclitean brain: Perspectivism and simplification in models of the motor cortex, *in* M. Massimi & C. D. McCoy, eds, 'Understanding Perspectivism: Scientific Challenges and Methodological Prospects', Routledge.

Coelho Mollo, D. (2015), 'Being clear on content', *Philosophia* **43**(3), 687–699.

Coelho Mollo, D. (2020), 'Content pragmatism defended', *Topoi* **39**, 103–113.

Cummins, R. C. (1989), *Meaning and Mental Representation*, MIT Press.

deCharms, R. C. & Zador, A. (2000), 'Neural representation and the cortical code', *Annual Review of Neuroscience* **23**, 613–647.

Dennett, D. C. (1978), *Brainstorms: Philosophical Essays on Mind and Psychology*, Bradford Books.

Dennett, D. C. (1991), 'Real patterns', *The Journal of Philosophy* **88**(1), 27–51.

Dretske, F. (1981), *Knowledge and the Flow of Information*, Basil Blackwell.

Dretske, F. (1986), Misrepresentation, *in* R. Bogdan, ed., 'Belief: Form, Content, and Function', Oxford University Press, pp. 17–36.

Egan, F. (2014), 'How to think about mental content', *Philosophical Studies* **170**, 115–135.

Fodor, J. A. (1987), *Psychosemantics*, MIT Press.

Fodor, J. A. (1990), *A Theory of Content and Other Essays*, MIT Press.

Frigg, R. (2010), 'Models and fiction', *Synthese* **172**, 251–268.

Godfrey-Smith, P. (2006), Mental representation, naturalism, and teleosemantics, *in* D. Papineau & G. Macdonald, eds, 'Teleosemantics: New Philosophical Essays', Clarendon Press.

Haugeland, J. (1990), 'The Intentionality All-Stars', *Philosophical Perspectives* **4**, 383–427.

Hutto, D. D. & Myin, E. (2013), *Radicalizing Enactivism: Basic minds without content*, The MIT Press.

Kaplan, D. M., ed. (2017), *Explanation and Integration in Mind and Brain Science*, Oxford University Press.

Kriegeskorte, N. & Douglas, P. K. (2018), 'Cognitive computational neuroscience', *Nature Neuroscience* **21**, 1148–1160.

Millikan, R. G. (2017), *Beyond Concepts: Unicepts, Language, and Natural Information*, Oxford University Press.

Mitchell, S. D. (2003), *Biological Complexity and Integrative Pluralism*, Cambridge University Press.

Morrison, M. (2015), *Reconstructing Reality: Models, Mathematics, and Simulations*, Oxford University Press.

Moser, E. I., Kropff, E. & Moser, M.-B. (2008), 'Place cells, grid cells, and the brain's spatial representation system', *Annual Review of Neuroscience* **31**, 69–89.

Neander, K. (2017), *A Mark of the Mental: in defense of informational semantics*, The MIT Press.

Potochnik, A. (2017), *Idealization and the Aims of Science*, University of Chicago Press.

Ramsey, W. M. (2007), *Representation Reconsidered*, Cambridge University Press.

Rescorla, M. (2009), 'Predication and cartographic representation', *Synthese* **169**, 175–200.

Rupert, R. D. (2018), 'Representation and mental representation', *Philosophical Explorations* **21**(2), 204–225.

Ryder, D. (2009), Problems of representation II: naturalising content, *in* F. Garzon & J. Symons, eds, 'The Routledge Companion to the Philosophy of Psychology', Routledge.

Shea, N. (2018), *Representation in Cognitive Science*, Oxford University Press.

Sprevak, M. (2013), 'Fictionalism about neural representations', *The Monist* **96**, 539–560.

Swoyer, C. (1991), 'Structural representation and surrogative reasoning', *Synthese* **87**, 449–508.

Thompson, E. (2007), *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*, Harvard University Press.

Toon, A. (2016), 'Fictionalism and the folk', *The Monist* **99**, 280–95.

Van Gelder, T. (1995), 'What might cognition be, if not computation?', *The Journal of Philosophy* **92**(7), 345–381.

Weisberg, M. (2013), *Simulation and Similarity: using models to understand the world*, Oxford University Press.

Weiskopf, D. A. (2017), The explanatory autonomy of cognitive models, *in* D. M. Kaplan, ed., 'Explanation and Integration in Mind and Brain Science', Oxford University Press.