Phenom Cogn Sci (2014) 13:257–274
DOI 10.1007/s11097-014-9350-6

Neural representationalism, the Hard Problem of Content and vitiated verdicts. A reply to Hutto & Myin (2013)

Matteo Colombo

Published online: 8 February 2014

Abstract Colombo's (*Phenomenology and the Cognitive Sciences*, 2013) plea for neural representationalism is the focus of a recent contribution to *Phenomenology and Cognitive Science* by Daniel D. Hutto and Erik Myin. In that paper, Hutto and Myin have tried to show that my arguments fail badly. Here, I want to respond to their critique clarifying the type of neural representationalism put forward in my (*Phenomenology and the Cognitive Sciences*, 2013) piece, and to take the opportunity to make a few remarks of general interest concerning what Hutto and Myin have dubbed "the Hard Problem of Content."

Keywords Neural representationalism · Hard problem of content · Un-metaphysical cognitive science

> "Confusing metaphysics with epistemology (and/or semantics) is the defining philosophical disease of our time, and it seems that the psychologists have caught it too"
> J. A. Fodor

1 Introduction

Hutto and Myin's (2013a) (henceforth H&M) take issue with the variety of representationalism put forward by Colombo (2013),[1] *viz*. neural representationalism. Here, I want to respond to H&M's critique and to take the opportunity to make a few, more general remarks concerning what Hutto and Myin have dubbed "the Hard Problem of

[1]My legal advisor Chiara reminded me that the legal sense of a plea is different from the colloquial sense. Legally, a plea is neither an earnest entreaty nor an apology. H&M over-trade on the colloquial sense of a 'plea,' which might give the mistaken impression that representations have been (or are about to be) convicted to perpetual philosophical and scientific exile.

M. Colombo (✉)
Tilburg Center for Logic, General Ethics, and Philosophy of Science (TiLPS), Tilburg University,
P.O. Box 90153, 5000 LE Tilburg, The Netherlands
e-mail: m.colombo@uvt.nl

Content." After rehearsing Colombo's (2013) aim, nature, and scope, and clearing up some possible misunderstandings of my arguments (Section 2), I shall make a few, more general, remarks on a misconception on which H&M's reaction relies (Section 3).

The misconception is captured by the claim that "unless Colombo—and others with similar ambitions—can explain how the Hard Problem of Content is to be overcome it looks as if their game is up. Without a solution to the Hard Problem of Content Colombo's plea for neural representations must be rejected" (H&M). This claim is misconceived because it is formulated on the basis of faulty understanding and/or disregard for the actual nature and scope of its target, *viz*. Colombo's (2013). More interestingly and generally, the claim suggests that unless the Hard Problem of Content is solved, *any* type of argument in support of *any* type of epistemological or methodological formulation (I shall not make a distinction here) of representationalism must be rejected. I shall argue that this more general claim is false. I shall begin to develop a broadly pragmatist, pluralist, and un-metaphysical rationale for positing neural representations that may provide a legitimate way of avoiding having to face up to the Hard Problem of Content. *Even if* the Hard Problem of Content hits the mark in revealing difficulties for those who want to connect some pragmatist defence of representationalism with any kind of realism, or with some other metaphysical tack on representations, purely un-metaphysical approaches will be left untouched. In the Conclusion, I shall summarize the take-home message of my argument. As it will be clear, the contribution of the present paper to existing literature is more the statement of an agenda for future work than the proposal of a detailed account of how the mind, whatever it is really like, should be studied and understood.

2 Colombo's (2013) neural representationalism. Aim, nature, and scope

Why would we want to argue about representationalism? I stated explicitly right at the beginning of my (2013) paper that, in articulating some arguments in support of a variety of representationalism, I intended to "shed light on the explanatory relationship between beliefs/preferences and norm compliance." The reasons for pursuing this project were twofold. First, establishing a link between belief/preference explanations of social norm compliance and some form of representationalism may advance our understanding of human moral/social psychology. Second, and more specifically, elucidating this link in terms of representations may advance our understanding of what types of manipulations and control on moral/social behaviour can be afforded by the explanatory relationship between certain beliefs and preferences, and social norms.

My project was carried out in three steps in that paper. First, I began by explaining that many prominent accounts of social norm compliance are cashed out in terms of beliefs and preferences, and that, in these accounts, the typical explanatory pattern for norm compliance is that people comply with social norms because they possess the right kinds of beliefs (or expectations) and preferences. However, these accounts do not specify what it can take to have beliefs and preferences such that they explain nom compliance. This makes it unclear whether understanding the appeal to beliefs and preferences in explanations of social norm compliance as an appeal to some sort of representations would give us any specific explanatory purchase, or, rather,

would be always detrimental to our epistemic interests, pragmatic goals, or methodological concerns.

With respect to this first step, it is worth clarifying that it is mistaken to say—as H&M do—that many working on social norms *suppose* or *assume* that beliefs are essential to make "social practices so much as possible." In the literature on social norms, this is neither a supposition nor an assumption, since it is not taken for granted that beliefs and norm compliance stand in any specific relationship. Claims about this relationship, about its nature, and about its bearing on rationality and on adequate explanation of social norms, are typically produced and justified through a process of *rational reconstruction*, which aims at capturing some essential features of social norms and at distinguishing social norms from moral norms, habits, conventions, and so on (cf. Bicchieri and Muldoon 2011). Rational reconstructions "specify in which sense one may say that norms are rational, or compliance with a norm is rational" (Bicchieri 2006, pp. 10–11). "A good rational reconstruction […] will also provide its own constraints: if we have a belief/desire model of norms, we must specify how behavior will change if beliefs change, and be able to make testable predictions" (Bicchieri and Muldoon 2011). So, it is mistaken to say that the literature on social norms endorses an "intellectualist assumption," *viz*. "the general edict that any and all intelligent activity involves having beliefs and desires" (H&M).[2]

The second step in my (2013) paper was the formulation of two types of approaches to what it is to believe and to prefer, *viz*.: dispositionalism and representationalism. Dispositionalism was construed as comprising an ontological and an epistemological thesis. Dispositionalism formulated as an ontological thesis claims that behavioural dispositions are fundamental to having beliefs and preferences, and that representations are "of only incidental relevance to the question of whether a being is properly described as believing" or preferring (Schwitzgebel 2006/2010). "For someone to believe some proposition P is for that person to possess one or more particular behavioral dispositions pertaining to P" (Schwitzgebel 2006/2010). Dispositionalism formulated as an epistemological thesis claims that invoking behavioural dispositions is often sufficient in order to make good sense of belief/preference explanations of behaviour, and that adequate belief/preference explanations of behaviour do not require positing or appealing to representations of any stripe.

The form of representationalism that I defended was called "neural representationalism." This view was construed as comprising two distinct claims: one ontological and one epistemological (see Chemero 2000 for why (anti)representationalism should be

[2] Relatedly, H&M write: "In line with received thinking in much economic and game theory, he [i.e. Colombo] tells us: 'Social norms are social… because we prefer to comply with them only if we believe that most members of our society will do the same.'"

This quote is importantly incomplete. The complete quote from my (2013) paper is: "Social norms are social, for Bicchieri, because we prefer to comply with them only if we believe that most members of our society will do the same, *and* we believe that most members of our society expect us to follow that norm and may sanction us with a reward or punishment depending on our choice to follow or violate the norm." Neglecting to take account of the full quote from my paper, H&M induce readers to incorrectly believe that "for those persuaded of this idea, any purported instance of social norm compliance will only really count as such if the agent in question harbours beliefs with the right contents – beliefs about what others are likely to do." (H&M). The claim is incorrect because those are not the "right contents" for those persuaded of that idea.

divided into an ontological and an epistemological claim). According to the first, ontological claim, what is essential to have beliefs and preferences is to have certain neural representations. According to the other, epistemological claim, neural representations are often necessary to adequately explain some cognitive phenomena and behaviour, whatever the mind really is. I was explicit that the latter, epistemological claim was the focus of my contribution.

After having provided one possible characterization of the concept of a neural representation, and after having suggested how beliefs and preferences can be understood in terms of this concept, the third and final step was to lay out three arguments—more in a moment on these arguments and on the reason for offering a characterization of the concept of a neural representation—for why the appeal to beliefs and preferences in explanations of some central cases of norm compliance should be understood as an appeal to neural representations. If any of those three arguments is persuasive, then there is reason to believe that neural representationalism yields some specific explanatory fruit with respect to social norm compliance. If there is reason to believe that neural representationalism yields some specific explanatory fruit with respect to social norm compliance, then some light has been shed on the explanatory relationship between belief/preference and social norm compliance.

The nature of the contribution in my (2013) paper should be sufficiently clear already. I was concerned with neural representationalism as an epistemological thesis "about the explanatory utility of invoking neural representations to explain certain cognitive phenomena and behaviour," in particular to explain some paradigmatic cases of social norm compliance.

I did not put forward any metaphysical argument. No claim was made about the actual existence of neural representations. I did not want to convince readers that we should believe in neural representations. Rather, I wanted to provide reasons for why we should believe that it is sometimes the case that understanding belief/preference explanation of certain behaviour and cognitive phenomena by positing neural representations bears some good explanatory fruit.

If metaphysics was not the aim, what was then the point of laying out a characterization of what neural representations could be? The point was to offer an unmetaphysically committed *definition* that could help make the use of the term 'neural representation' more easily understandable in the three arguments for the explanatory utility of neural representationalism. According to that definition, neural representations should be understood as encoding-decoding mappings between two alphabets constituting a neural code. While this definition coheres with the general idea of a representation as something that stands in for something else and that can be used to guide behaviour, it imposes some constraints on the proper assessment of my (2013) representationalist, epistemological, thesis. My claims about what neural representations could be were intended as claims about a concept embedded in scientific theorizing as actually practiced, which—let me add—may also receive a precise mathematical formulation (e.g. deCharms and Zador 2000; Pouget, Dayan, and Zemel 2003). Thus, the assessment of this definition of the concept of a neural representation ought to be based on its overall coherence with actual scientific theorizing, as well as on its usefulness for particular epistemological or methodological purposes in particular contexts, rather than in terms of its adequacy in capturing what neural representations really are. As I shall elucidate in the next section, other representationalist,

un-metaphysical, projects may well adopt the same type of strategy for treating such a theoretical posit as 'neural representation'.[3]

Because I was not fully clear on the point of providing a characterization of a concept of neural representation in that paper, H&M could assume that there was a metaphysical account backing my epistemological claims. If this were so, then their criticisms would be on target. However, if I was uninterested in defending a realistic, metaphysically robust, theory of neural representations—as I was indeed—and I was offering an un-metaphysical definition of a concept of neural representation, then H&M's critique is off the mark. In an attempt to move forward the type of debate between Colombo (2013) and H&M, below I hope to drive home a more interesting and general point, that is: the assumption that representationalist epistemological/methodological theses always need metaphysical backing from the existence of representations is misconceived.

Let me first elucidate what the scope of my neural representationalism was. My claim was that, for some cases of social norm compliance, belief/preference explanations of such cases should be understood in terms of neural representations. Put differently, the claim was that, at least sometimes, positing neural representations is explanatory fruitful with respect to some cognitive phenomena and behaviour. No claim was made as to whether all cases of social norm compliance always involve neural representations. No claim was made as to whether all explanations of social norm compliance should be understood as positing neural representations. Neither was it claimed that positing neural representations is sufficient to adequately explain all cases of social norm compliance, nor that neural representationalism provides the best explanation for all cases of norm compliance.

It is not unambiguous what H&M take the scope of my contribution to be. In some places, they correctly understand the scope of my argument as restricted to "central cases of social norm compliance." However, in some other places, they omit the adverbial qualifiers and determiners that are necessary to accurately characterise what I sought to establish. Thus, they appear to suggest that my argument carries over to our "capacity to adhere to social norms" (H&M), or to all instances of social norm compliance. This is apparent when H&M write: "What would follow if Colombo turned out to be right that neural representations are ultimately needed to explain central cases of social norm compliance? A pivotal premise in Colombo's argument is that social norm compliance is a representation-hungry phenomenon. Hence, it is safe to assume that if his argument works its conclusion will generalize to some significant extent. Other so-called 'representation-hungry' phenomena can expect to get the same treatment" (H&M). In fact, the premise in that argument is that *some* cases of social

[3] This type of strategy is not novel. It is very much in the same spirit as the un-metaphysical, methodological treatment of theoretical concepts put forward by Rudolf Carnap. In his seminal "The Methodological Character of Theoretical Concepts" (1956), writes Carnap: "We have considered some of the kinds of entities referred to in mathematics, physics, psychology, and the social sciences and have indicated that they belong to the [purely mathematical] domain D. However, I wish to emphasize here that this talk about the admission of this or that kind of entity as values of variables in [the theoretical language] L_T is only a way of speaking intended to make the use of L_T, and especially the use of quantified variables in L_T, more easily understandable. Therefore, the explanations just given must not be understood as implying that those who accept and use a language are thereby committed to certain "ontological" doctrines in the traditional metaphysical sense. The usual ontological questions about the "*reality*" (in an alleged metaphysical sense) of numbers, classes, space-time points, bodies, minds, etc., are pseudo-questions without cognitive content" (Carnap 1956, pp. 44–45).

norm compliance are properly characterised as consisting in behaviour in representation-hungry problem domains. It was not claimed that social norm compliance is a representation-hungry phenomenon without qualification. If that argument succeeds, then it is only safe to assume that the conclusions would generalize in a rather qualified way, only to those particular case studies that are properly characterised as representation-hungry.

Similarly, H&M have not been fully accurate in reconstructing Colombo's (2013) arguments. Three claims were put forward: first, some instances of norm compliance take place in "representation-hungry problem domains"; second, neural representationalism often facilitates manipulation and control of some behaviour and cognitive phenomena; third, neural representationalism yields non-trivial understanding of current explanatory practice in some areas of cognitive science.

Much of H&M's counter examines what H&M take to be the first of the three arguments. That argument relied on Clark and Toribio's (1994) notion of a "representation-hungry" problem domain. The argument is as follows: internal stands-in or representations give us "unique explanatory" leverage about some instances of norm compliance because there is reason to believe that some cases of social norm compliance consist in "representation-hungry" problem domains. I took for granted the premise that "[i]nternal representations give us unique explanatory leverage regarding agents' behaviour in 'representational-hungry' problem domains" noting that Clark and Toribio (1994) provided good grounds to accept it.

It is worth recalling that two main points are made by Clark and Toribio (1994). First, in assessing a given representationalist thesis, they recommend to not conflate different notions of representation, to not conflate 'representation' specifically and classically understood as a symbol string in a compositional declarative code with a more general notion of representation (see Haugeland 1991 for a nice discussion of several distinct genera of representation). For "there is a rich continuum of degrees and types of representationality," to which we may justifiably appeal in explanations of some cognitive phenomena but not others (Clark and Toribio 1994, p. 401).

Second, and more relevant here, Clark and Toribio (1994) argue that there is *prima facie* reason to hold that appealing to at least some 'modest notion of representation'[4] gives us unique explanatory leverage for cases in which the behaviour or the cognitive phenomena displayed by a system involve some sensitivity to distal, non-existent, or abstract properties. In such cases, the system cannot rely on a direct coupling to the relevant ambient environmental feature, which would allow for adaptive coordination with the environment. In such cases, adequate explanations of behaviour and cognitive phenomena would require an appeal to specific resources and operations, which would justify invoking some (at least modest) concept of representation. Such resources and operations would consist in the filtering, re-coding, compression and dilation of available information in the sensory input array through a cascade of layers of

[4] Clark and Toribio characterise 'modest internal representations' as "internal information bearing states which capture regularities which are not available in the simple surface statistics of the input arrays, but instead emerge only as a result of the subsequent filtering and transformation of such signals" (Clark and Toribio 1994, p. 421).

processes so as to enable the system to track distal, non-existent, or abstract properties, and in the subsequent use of the re-coded information so as to guide behaviour.

Clark and Toribio's point is *not* that "it is simply inconceivable that object recognition, counterfactual reasoning and selective response to rather abstract kinds of features might all succumb to some unexpected, representation-free, kind of explanation" (Clark and Toribio 1994, p. 421). Rather, their point is that if a system is to deal with certain kinds of problem-domains, then an adequate explanation of the system's behaviour or of the cognitive phenomena it displays will *prima facie* need to appeal to some concept of representation, given the resources and operations that are likely to be recruited in those domains.

In my (2013) paper, I argued for the premise that some central cases of norm compliance are rightly characterised as representation hungry. And so, appealing to some concept of a representation would provide us with unique explanatory leverage with respect to those cases. Subsequently, I considered some objections to my argument by examining Dreyfus's (2002a, 2002b) diagnosis of how central cases of social norm compliance are rightly characterised. I provided a number of reasons for why Dreyfus's diagnosis should be resisted.

In analysing this first argument from Colombo (2013), H&M miss the target. To be on target, H&M's critique should have taken the care to explain why none of the cases I examined is rightly characterised as representation-hungry. Instead, H&M argue against the notion of a "representation-hungry" problem domain, and its relationship to explanation of behaviour and cognitive phenomena. However, H&M do not examine the relevant arguments, those put forward by Clark and Toribio (1994), which were rehearsed above.

Interestingly, neither Hutto and Myin's "Radicalizing Enactivism" (2013b) examines Clark and Toribio's (1994) arguments for why appealing to some, non-necessarily classicist, notion of representation gives us unique explanatory leverage with respect to representation-hungry problem-domains; in fact, Hutto and Myin needed not to counter those arguments in their book, since the thesis they set out to defend is that "not all mentality requires individuals to construct representations of their worlds" (Hutto and Myin 2013b, p. 5), which is consistent with Clark and Toribio's (1994) claims, and—let me add—with my claims too.

The second argument put forward by Colombo (2013) focused "on requests for explanation that aim for manipulation and control of social norm compliance." Insofar as understanding explanations of social norm compliance in terms of a certain concept is particularly fit to afford information useful for identifying specific kinds of interventions, the concept bears some epistemic/explanatory fruit. Pitting dispositionalism against neural representationalism, I argued that at least some belief/preference explanations of a given behaviour understood in terms of neural representations are fitter than explanations understood in a dispositionalist framework to afford reliable control and manipulation on the behaviour of interest. In the light of recent advances in manipulative neuroscience (Kawato 2008; see also e.g. Nicolelis 2001; Nicolelis and Lebedev 2009), it was explained that neural representationalism helps to identify possible interventions not only on particular neural structures or populations of neurons, but also on the particular signals carried by these structures. So, it is mistaken to say that "[a]ll the benefits for potential interventions pointed at by Colombo will be shared by

any structural account that targets a deeper level, including a non-contentful, non-representational one" (H&M).

My point here is not that it is *inconceivable* that an understanding of some target belief/preference explanation in non-representationalist terms—or in terms of some other concept of representation—*can* afford reliable control and manipulation on the behaviour of interest. The point that bears emphasis is rather that, *currently*, in the light of explanatory practice in some prominent areas of cognitive science, a certain concept of neural representation is particularly fit to afford information useful for identifying certain kinds of interventions. The questions on which the case for the kind of neural representationalism put forward by Colombo (2013) should rest are: In the light of scientific practice, to what extent does the use of a given concept help to understand what types of interventions are most appropriate in a certain context, in order to achieve certain pragmatic goals? And, generally, what roles does such a concept play in a particular scientific practice?

Indeed, my third argument was that "in order to understand current explanatory practice in cognitive science, invoking neural representations is often necessary." To make this point more precise, it may be examined how probabilistic and/or reinforcement learning models can be used to account for certain cognitive phenomena and behaviour (cf., Dayan 2008; Friston 2005; Knill and Pouget 2004; Tenenbaum et al., 2011; see Colombo 2014 for a proposal about how social norm compliance might be explained within a Bayesian-RL neurocomputational framework). The basic insight of many of these models is that several cognitive phenomena and behaviour can be explained by computations, carried out by cognitive systems, of differences (or errors) between the way the world is modelled as being, and the way the world actually is (Shea 2012 provides a thorough defence of the thesis that reward-prediction errors posited by some RL models have genuine, real, meta-representational contents). For example, according to some broadly Bayesian accounts, brains make inferences about the (hidden) causes of their sensory inputs in order to enable perception, action, and adaptive behaviour (e.g., Clark 2013). If brains are understood as making inferences about the (hidden) distal causes of their sensory inputs, then brain can justifiably be understood as possessing some model of the causal/statistical relationships among (hidden) states of the world—not directly present to the cognitive system—that cause sensory inputs. Given sensory inputs, brains would make inferences over such models so as to generate cognitive phenomena and behaviour. The relationship between causal models and external causal structures can be more easily understood by appealing to the notion of probabilistic/structural representation: the models stand in for causal, probabilistic structures in the world, thereby mediating between external environmental features, internal processes, and behaviour.

Some concept of representation is useful not only to elucidate how probabilistic and reinforcement learning models can be used to account for cognitive phenomena and behaviour, but also to illuminate some of the reasons for taking these general approaches to studying the mind. For example, uncertainty[5] is a fundamental and unavoidable feature of cognitive systems' interaction with the world. Given this feature,

[5] 'Uncertainty' here refers to a lack of information for a cognitive system, which is due either to noise—that is, to random disturbances corrupting sensory signals and cognitive processing—or to the underdetermination of percepts as well as of other cognitive states by input data.

one reason for building a model of cognitive phenomena within a probabilistic framework is that such a framework provides one effective, flexible, language to represent and deal with uncertainty. If the human cognitive system has to handle uncertainty in order to produce at least certain cognitive phenomena and adaptive behaviours, then the system can be modelled as representing "information probabilistically, by coding and computing with probability density functions, or approximations to probability density functions," which mediate the production of those phenomena and behaviours (Knill and Pouget 2004, p. 713).

"Should we base our faith in neural representations on the fact that cognitive scientists use models that invoke them?" ask H&M. Their answer is: "Well, we should if the theories those models figure in are true." This answer does not address my point. For my point is that neural representationalism illuminates current explanatory practice in prominent fields of contemporary cognitive science. In the light of a concept of neural representation, it is easier to understand how certain models can be used to account for cognitive phenomena and behaviour, and it becomes clearer why a certain approach to study cognition is currently pursued.

It should be pointed out that two distinct issues seem to be conflated in H&M's question-and-answer above. First issue: does some concept of representation illuminate some aspects of explanatory practice in cognitive science? Second issue: do models that invoke neural representations provide some insight into cognition and behaviour? Colombo's (2013) third argument does not address the second issue; it concentrated on the first issue only. If invoking neural representations provides insight into why a certain approach is pursued for the study of cognition, or how some current models are used to provide an explanation of some cognitive phenomena and behaviour, then there is reason in support of neural representationalism. So, H&M should have showed that, for all Bayesian and RL models, neural representationalism provides no insight with respect to why or how these models are built, and used to account for cognitive phenomena and behaviour.

The idea that my assessment "blatantly, and without warrant, fails to take stock the new, non-representational developments afoot in the cognitive sciences" is irrelevant. For—as it should be clear by now—my purpose is not to persuade the reader that neural representations are necessary and sufficient to explain all (or most) cognitive phenomena. My position allows for the possibility that, for some cognitive phenomena and behaviour, neural representations do not bear much explanatory fruit, or bear no explanatory fruit at all. Although neural representations are no panacea, invoking them is often not only explanatorily useful, but also congenial to illuminate the explanatory practice in some fields of contemporary cognitive science.

Once the aim, nature, and scope of Colombo's (2013) are clear, it will also become clear that H&M's "final verdict" is multiply vitiated. For it is based on an inaccurate rendering of their alleged opponent's claims. The "verdict" should then be rejected.

3 Should representationalists always face up to the Hard Problem of Content?

The "Hard Problem of Content," as H&M dub it (see also Hutto and Myin 2013b, Ch. 4), is in essence the good old problem of naturalizing representational content

(a.k.a. naturalizing intentionality[6]). One way to understand content is in terms of the set of properties ascribed to something by its representation. Content is what a representation tells us about its target. The problem of naturalizing representational content consists in providing a satisfactory answer to two questions: 1) What is the nature of mental representations? 2) How can mental representations be about things? These questions concern the metaphysics of representations. The first question asks what kind of entity a mental representation is. Possible (non-mutually exclusive) answers include: mental representations are neurobiological states; they are some sorts of processes; they are classical computational symbols; they are activation vectors; they are images, and so on. The second question asks about the semantic properties of mental representations. It asks what makes it the case that a mental representation has the content it has. In order to address this question, one needs to provide an account of how the set of properties ascribed to something by its representation is determined.

H&M appear to suggest that unless the Hard Problem of Content is solved, any type of epistemological argument in favour of representationalism must be rejected. They write: "unless Colombo—and others with similar ambitions—can explain how the Hard Problem of Content is to be overcome it looks as if their game is up. Without a solution to the Hard Problem of Content Colombo's plea for neural representations must be rejected" (H&M).

Two claims should be distinguished here. First claim: for all type of representationalism, if one does not reject representationalism, then she should face up to the Hard Problem of Content. She should provide a convincing account of what makes it the case that the representations she posits, relies upon, accepts, believes to exist, or, more plainly, does not reject, have the contents they do. Second claim: for all type of argument in support of some form of epistemological representationalism, if one's argument has hope to succeed, then she should face up to the Hard Problem of Content. The idea is that the prospects of success of any argument for epistemological representationalism are conditional upon a solution to the problem of naturalizing representational content.

Both claims are misconceived. For they are based on the faulty idea that one should face up to a metaphysical problem when she may be arguing just for some form of epistemological view or she may be interested in elucidating some epistemological/methodological issues surrounding a certain theoretical posit. If H&M were right, there could be no room for any un-metaphysical representationalism, or for arguments in support of epistemological versions of representationalism.

[6] More accurately, the problem of intentionality asks how mental items such as thoughts, beliefs, and desires can be directed towards, or be about, other specific items. The problem of representation asks how certain kinds of items, viz. representations, can represent, can be directed towards, or be about, other items. The concepts of *intentionality* and *representation* are distinct, and in fact the notion of representation can be used as a means to address the problem of intentionality. Yet, the problem of representation and the problem of intentionality are often taken to be identical. Just to give two examples, Fodor (1987, p. xi) writes: "It appears increasingly that the main joint business of the philosophy of language and the philosophy of mind is the problem of representation itself: the metaphysical question of the place of meaning in the world order." In an introduction to contemporary philosophy of mind, Crane (2003, p. 30) writes: "those mental states which exhibit intentionality –those which represent – are sometimes therefore called 'intentional states'." Interestingly, Fodor himself is in fact aware of the distinction when he observes that his own representationalism "doesn't, of course, *solve* the problem of intentionality; it merely replaces it with the *unsolved* problem of representation" (1996, p. 260).

Furthermore, those claims badly characterise the scientific community's own understanding of when a theoretical posit can be justifiably invoked and relied upon. If H&M were right, scientists could not invoke or use a theoretical posit quite successfully in default of a metaphysically thorough account of what that posit picks out in the world. But, in fact, many theoretical posits are invoked and relied upon quite successfully by scientists, even though there is currently no metaphysically thorough or convincing account of those posits. Among such posits we find: *gene*, *phoneme*, *quark*, *string*, and *space-time*.

Consider the alleged parallelism between a *gene* and a *representation*. Some may believe that the parallelism is unjustified. For, unlike the notion of a representation, there is an accepted answer about how to account for the central feature of a gene, namely that it is a crucial mediator of heredity. Instead, in the case of representation—as H&M point out—there is no accepted or satisfactory answer of how to account for the central feature of a representation, namely that it is about, or stands in for some other thing. This seems to put *representation* in a different league than *gene*.

The problem with this line of reasoning is that it oversells the consensus around the notion of a gene, which is at least as controversial as the notion of a representation. No scientific or philosophical consensus has been reached on how to best understand what a gene is and what genes do (cf., Waters 2013, Sec. 4–5). Rheinberger and Müller-Wille (2010) capture this lack of consensus when they write: "'There can be little doubt,' philosopher and biochemist Lenny Moss claimed in 2003, 'that the idea of 'the gene' has been the central organizing theme of twentieth century biology' […]. And yet it is clear that the science of genetics never provided one generally accepted definition of the gene. More than a hundred years of genetic research have rather resulted in the proliferation of a variety of gene concepts, which sometimes complement, sometimes contradict each other […]. Today, along with the completion of the human genome sequence and the beginning of what is being called the era of post-genomics, genetics is again experiencing a time of conceptual change, voices even being raised to abandon the concept of the gene altogether in favor of new terminologies."

Gene and representation are notions more similar than what some might suspect. They are central theoretical constructs of biology and cognitive science, respectively. Analogously to the notion of representation, the notion of a gene is much different in biology than it was 100 years ago, and it is a matter of current lively debate what a gene is and what genes do. Nonetheless, most scientists do not seem to worry about this lack of consensus; they continue to talk about genes in a pluralist, contextual, and pragmatically-grounded way; and the notion of a gene continues to be central to their explanatory and epistemic practices.

When confronted with such theoretical constructs as gene or representation, philosophers of science often try to contribute to the conceptual foundations of a discipline. They aim at illuminating the theoretical posit in question by examining how the relevant scientific community think about, learn, and reason with the target notion, by drawing conceptual links between the target notion and other notions used by the relevant science, and by considering the roles that such notion plays in explanation, prediction, control, as well as in other epistemic or pragmatic scientific projects. Such epistemological/methodological projects do not obviously bear on metaphysics, and metaphysics does not obviously bear on them. These projects may establish that a target

theoretical posit is explanatorily and pragmatically inert. But, *at most*, this epistemological conclusion may serve as evidence for the separate claims that the theoretical posit in question does not pick out anything at all in the world, or that the theoretical posit, as it figures in scientific practice, is significantly different from some homonymous notion figuring in common, non-scientific discourse (cf. Cummins 1989 and Ramsey 2007 for two nice examples of this type of project with respect to the notion of representation).

If one's project is to develop and support some form of epistemological representationalism, then she may do so by showing that that form of representationalism illuminates how cognitive scientists think about, learn about, and reason with various representational posits and about the roles that such posits play in explanation, and in other scientific practices. However, for this project to be successful, she need not commit herself to any fundamental metaphysical account of representation. For successfully carrying out such types of epistemological/methodological projects, one can provide *definitions* of a concept of representation in much the same vein as explained in the previous section, which is to be judged by their usefulness in providing understanding of cognitive phenomena and behaviour, in illuminating explanatory practice in cognitive science, in drawing fruitful connections between different notions, in affording compact, intelligible descriptions of cognitive phenomena, and so forth. From this perspective, "the question should not be discussed in the form: "Are theoretical entities real?" but rather in the form: "Shall we prefer a language of physics (and of science in general) that contains theoretical terms, or a language without such terms?" From this point of view the question becomes one of preference and practical decision" (Carnap 1974, p. 256).

For those who care also about metaphysics, a given epistemological/methodological account can be supplemented by facing up to the Hard Problem of Content, providing some plausible explanation about the nature and semantics of mental representations. Yet, it is misconceived to demand that any type of project concerned with supporting representationalism should attempt to provide also an account of the metaphysical foundations of representation. Some version of representationalism can be justifiably accepted based on legit arguments, based on arguments aimed at showing that some notion of representation bears explanatory fruit, plays some central role in scientific practice, or is pragmatically valuable, without any obligation to face up to the Hard Problem of Content.

This point—the point that some un-metaphysical version of a pragmatically and epistemologically grounded representationalism can be legitimately defended—is borne out by examining some central contributions to the philosophy of science. Two of the most widespread and important notions underlying many scientific and philosophical projects are: *causation* and *mechanism*. Consider causation. It is controversial what the nature of causal relations is and what the nature of causal *relata* is. Apparently the metaphysical foundations of causation are not stable. Nonetheless, scientists continue to fruitfully posit and rely on the notion of causal relationship in their explanatory projects, and some philosophers of science, who are less concerned with metaphysical issues than with epistemological/methodological ones, develop and argue for un-metaphysical accounts of causal explanation, providing arguments in support of some account of causal explanation that are neutral with respect to fundamental metaphysical debates.

Woodward (2003), for example, developed an interventionist account of causal explanation. The perspective of Woodward's (2003) account can be described "as that of a modeler: pragmatic, piece-meal, and anti-foundational" (Woodward 2008, p. 195). Woodward's (2003) focus is not the metaphysics of causation: he urges that we define causal relationships simply as relationships that can be exploited for purposes of manipulation and control. While this characterization carries little metaphysical baggage, and so it allows one to remain neutral on metaphysical foundational issues about causation, the nature of Woodward's (2003) contribution is "*methodological*: [it aims to elucidate] how we think about, learn about, and reason with various causal notions and about their role in causal explanation" (Ibid.).

Woodward's (2003) account of causal explanation should be properly defended or attacked on purely un-metaphysical grounds. The superiority of this account over competitors may be shown in its success in advancing our understanding of several causal notions and of their roles in explanation, prediction, and description, in a way that is sensitive to the actual explanatory concerns and pragmatics goals of scientists. Similar un-metaphysical projects, which argue for the fruitfulness of certain accounts of causal explanation, have been taken up by several other researchers in the last decade or so (e.g., Pearl 2000; Spirtes, Glymour, and Scheines 2000). Whether these accounts lack adequate metaphysical foundations does not detract from their success; and lack of deep metaphysical foundations is not an argument against them. Rather—as Woodward (2003; 2008) argues—the un-metaphysical character of accounts such as his should be counted as one of their attractive features.

Consider another example. Many phenomena studied in the special sciences are often explained by specifying the mechanisms that bring them about. Moreover, "in many areas of science, explanations are said to be adequate to the extent, and only to the extent, that they describe the causal mechanisms that maintain, produce, or underlie the phenomenon to be explained, the explanandum phenomenon" (Kaplan and Craver 2011, p. 601). This is a central thesis of the mechanical account of scientific explanation. However, it is far from being uncontroversial what a mechanism is. And it is far from being settled whether adequate mechanistic explanation requires a commitment to some form of scientific realism (Colombo et al. 2014). This lack of agreement about metaphysical foundations notwithstanding, the notion of a mechanism is at the core of the mechanistic account of explanation, and it appears to be able to serve well cognitive scientists' interests for prediction, description, intervention and explanation. In order to satisfy such interests, it need be provided no deep metaphysical account of what mechanisms are. It often suffices to think of a mechanism as a set of entities and their associated causal activities spatio-temporally organized in such a way as to produce some phenomenon of interest. Spelling out the details of this characterisation is obviously important if one is concerned with the metaphysics of mechanisms. Then, one should develop an account of what kinds of entities feature in a mechanism, how they constitute the mechanism, what sort of thing is the production-relationship between a mechanism and its phenomenon, and so on. But the fate of a mechanistic account of scientific explanation will not depend entirely on the ability of philosophers to provide a metaphysically satisfying account of what mechanisms are.

It may be objected that while it is correct to say that the notions of causation and mechanism, even in a deflated understanding, are somewhat less than adequately defined but remain nevertheless exceptionally useful in scientific and everyday practice

alike, the trouble with neural representations is that, unlike the notions of causation and mechanism, they are positively defined as having properties that they may not have, if the Hard Problem of Content is not answered. Hence, the comparison between causation/mechanism and neural representation is illegitimate.

However, this complaint should be resisted, at least for three reasons. First, *causation* and *mechanism* are here used as examples where some notion may lack firm metaphysical foundations, and yet it is the object of epistemological/methodological accounts, or it is fruitfully invoked and relied upon for satisfying a number of epistemic, methodological or pragmatic interests and goals. As I pointed out, defensible, highly successful, un-metaphysical theories of causation or of mechanistic explanation can be given. The situation for representation and some versions of representationalism is parallel in this respect—or so I claim. The parallelism—let me be clear—is not that talk of mechanisms or causation in the sciences generally plays anything analogous to the explanatory roles that theoretical constructs such as neural representations or genes are meant to play in the sciences of the mind or in biology. The parallelism here is simply that, analogously to some un-metaphysical accounts of causation or of mechanistic explanation, the fate of a pragmatically and epistemologically grounded representationalism does not depend on the availability of a satisfying answer to metaphysical issues.

Second, *if* the notion of a neural representation is positively defined as having properties that—if the Hard Problem of Content is not answered—they may not have, the notions of mechanism or causation do not enjoy a metaphysically much better standing. Take *causality*. It has been common in philosophy to define the causal relation in terms of one of several other notions, including the notions of natural law, statistical correlation, counterfactual dependence, and physical process. For each of these notions, the difference between causally related and causally unrelated sequence is positively defined in terms of a specific metaphysical basis. According to process theories of causation, the metaphysical basis for causal connection is some sort of physical producing (cf. Dowe 2008). If the objections levelled against the notion of *physical producing* are not answered, then causal relations would be positively defined as relying on a metaphysical basis that they may not have. Or consider mechanism. Most accounts agree that there are two kinds of constituents of mechanisms. Specifically, some prominent account is committed to a dualistic ontology of entities and activities (Machamer et al. 2000), which has recently raised several perplexities. A number of problems have been identified for the concept of activities (e.g. Psillos 2004; Tabery 2004; Persson 2010; see Machamer 2004 on the metaphysics of activities). If these problems are not adequately answered, then mechanisms would be positively defined as having a kind of constituent that they may not have.

Third and finally, disputes over whether an explanation is really mechanistic or whether a system is really representational are largely terminological. The answers to such questions depend on what is meant by 'mechanism' and 'representation.' But there are no universally accepted definitions of either notion; and, plausibly, both 'mechanism' and 'representation' admit narrow and broad characterizations. Both mechanistic/non-mechanistic and representational/non-representational contrasts are better understood not in terms of a sharp dichotomy; but rather in terms of a graded, multi-dimensional continuum, according to which some systems are more mechanical or representational than others, along various dimensions (see Clark and Toribio 1994

for this idea with respect to representation, and Woodward 2013 for why mechanisms should be understood as graded). What is important here are to identify similarities and differences between systems, and to perspicuously characterise them in a way congenial to one's epistemic and pragmatic goals, interests and needs.

There is little reason to believe that unless the Hard Problem of Content is solved, un-metaphysical arguments in support of representationalism or epistemological versions of representationalism such as Colombo's (2013) must be rejected—as H&M appear to claim. A complete, satisfactory metaphysical representationalist account, one that successfully addresses the Hard Problem of Content, does not entail that our best models and explanations should only invoke representations. It may well be the case that some theoretical posit other than representation will provide us with some unique explanatory insight, by e.g. guiding some kinds of interventions on certain systems of interest, or by generating precise predictions with respect to a given phenomenon, or by simply affording more compact descriptions of phenomena and systems of interest that could facilitate scientists to communicate and share information.

Truth or existence is not a necessary condition for theoretical posits like representations to be legit epistemological/methodological tools. A successful version of epistemological representationalism does not entail that representations exist. Besides, let me note that there is reason to believe that, for at least some theoretical posits, truth or existence is not necessary for those posits to do good explanatory work (cf., Bokulich 2008, 2012; Frigg 2010; see Sprevak 2013 for a critical discussion of a fictionalist attitude towards neural representations). So, *even if* biological cognitive systems are not representational systems at all, some form of epistemological/methodological representationalism can still be successfully defended, and representation-talk can still be justifiably preserved.

In sum, I hope that the remarks in this section have contributed to untie some of the knots in our thinking about representations. A lack of solution to the Hard Problem of Content should not dissuade us to pursue either of two epistemological/methodological projects. The first project aims at contributing to the conceptual foundations of cognitive science, by illuminating the role of that the concept of representation plays in scientific practice. More relevant here, the second project aims at establishing some form of epistemological representationalism, according to which positing representations is often explanatory fruitful. I suspect—for the little that it is worth—that most cognitive scientists and philosophers of cognitive science who care about metaphysics are convinced that representational content can be naturalized in terms of some relational and/or causal properties of neurobiological states or processes. Whether an appropriate account of representational content can actually be given remains controversial. Even if it cannot be given, little hangs on the matter.

4 Conclusions

In a great many historical cases, scientific progress has been marked by breaking up with conceptual inertia that contributed to preserve the theoretical *status quo* of a discipline. H&M, at one point, claim that arguments I put forward in my (2013) paper

may well be used to try and save faulty theories from rejection and revision, so as to unhelpfully preserve the *status quo*. Shying away from facing up to the Hard Problem of Content might be seen as contributing to conceptual inertia, which may hinder progress in our understanding of brains and cognition.

Underlying the exchange between H&M and me, some readers might have the impression that there are two mutually exclusive visions of the prospects and goals of cognitive science. Each vision would claim that there is one correct account of cognition and any other competitor is incorrect; this vision would supposedly hang together with the idea that there is one, most fruitful theoretical framework within which cognition should be studied. Although rhetorical excess might suggest otherwise, this impression is unfounded. It is also unwelcome, as it may really hinder progress in our understanding of brains and cognition.

I see no reason to suppose that there is a unique correct theory of cognition or a single, most fruitful theoretical framework where we should study cognitive phenomena. There are myriads of interesting cognitive phenomena and behavioural regularities to explain, and there are several levels of description and analysis at which a given phenomenon can be studied. Given this multiplicity, one theory may well prove correct with respect to some phenomenon but not with respect to others. And studying a given phenomenon within different frameworks may yield explanatory fruit that could not be achieved if radicalism were to be pursued. Even if it were possible that a single conceptual framework could be used to explain all cognitive phenomena and behaviour, it would not follow that the best explanation for a given phenomenon of interest will always be found within that framework. It could still be possible—and probable indeed—that some alternative framework, some alternative notion or theoretical posit will yield uniquely good explanatory fruit with respect to that phenomenon for specific purposes.

Different kinds of representationalisms might be suited for different tasks; some will be specifically fit to explain different kinds of systems, others to illuminate different aspects of explanatory practices in cognitive science. Analogously, other non-representational views will be more suited for other, different explanatory tasks. Notions like *representation* or *computation* will change, and will need fine-tuning, but they do not need to be exiled from our conceptual arsenal.

Furthermore, as Shea (2013, p. 502) has thoughtfully suggested, "there may be no one true unified account of the nature of content. The metaphysics of content may be different in different kinds of representational system." So, there may well be no unified answer to the Hard Problem of Content.

The time is ripe for antirepresentationalists and representationalists alike to give up their radical ambitions. If there is a knot that need be untied in our thinking about brains and cognition, that knot has not to do with representationalism—neural representationalism or otherwise—or with the inability to facing up to the Hard Problem of Content. That knot is the unfruitful prejudice that there is *one* correct way to go about to study brains and cognition.

Acknowledgments I am grateful to two anonymous reviewers for this journal for their constructive comments and helpful suggestions. This work was supported by a grant from the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program "New Frameworks of Rationality" (SPP 1516). The usual disclaimers about any error or mistake in the paper apply.

References

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge: Cambridge University Press.

Bicchieri, C. & Muldoon, R. (2011). Social Norms, *The Stanford Encyclopedia of Philosophy* (Spring 2011 Edition), Edward N. Zalta (ed.), URL=http://plato.stanford.edu/archives/spr2011/entries/social-norms/

Bokulich, A. (2008). *Reexamining the quantum-classical relation: Beyond reductionism and pluralism*. Cambridge: Cambridge University Press.

Bokulich, A. (2012). Distinguishing explanatory from Non-explanatory fictions. *Philosophy of Science, 79*(5), 725–737.

Carnap, R. (1956). The methodological character of theoretical concepts. In H. Feigl & M. Scriven (Eds.), *Minnesota studies in the philosophy of science. The foundations of science and the concepts of psychology and psychoanalysis* (Vol. I, pp. 38–76). Minneapolis: University of Minnesota Press.

Carnap, R. (1974). *An introduction to the philosophy of science*. New York: Basic Books.

Chemero, A. (2000). Anti-representationalism and the dynamical stance. *Philosophy of Science, 67*, 625–647.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Science, 36*, 181–253.

Clark, A., & Toribio, J. (1994). Doing without representing? *Synthese, 101*, 401–431.

Colombo, M. (2013). Explaining social norm compliance. A plea for neural representations. *Phenomenology and the Cognitive Sciences* online first 2013. doi:10.1007/s11097-013-9296-0.

Colombo, M. (2014). Two neurocomputational building blocks of social norm compliance. *Biology and Philosophy, 29*(1), 71–88.

Colombo, M., Hartmann, S., & van Iersel, R. (2014). Models, Mechanisms, and Coherence. *The British Journal for Philosophy of Science* (in press).

Crane, T. (2003). *The mechanical mind* (2nd ed.). London: Routledge.

Cummins, R. (1989). *Meaning and mental representation*. Cambridge: Bradford Books/MIT Press.

Dayan, P. (2008). The role of value systems in decision making. In C. Engel & W. Singer (Eds.), *Better than conscious? decision making, the human mind, and implications for institutions* (pp. 51–57). Frankfurt: MIT Press.

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

Dowe, P. (2008). Causal Processes. *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), Edward N. Zalta (ed.), URL= http://plato.stanford.edu/archives/fall2008/entries/causation-process/

Fodor, J. (1987). *Psychosemantics*. Cambridge: MIT Press.

Fodor, J. (1996). Deconstructing Dennett's Darwin. *Mind and Language, 11*, 246–262.

Frigg, R. (2010). Fiction in Science. In J. Woods (Ed.), *Fictions and models: New essays* (pp. 247–287). Munich: Philosophia Verlag.

Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B, 360*, 815–836.

Haugeland, J. (1991). Representational Genera. In W. Ramsey, S. Stich, & D. Rumelhart (Eds.), *Philosophy and connectionist theory* (pp. 61–89). Hillsdale: Lawrence Erlbaum.

Hutto, D. D., & Myin, E. (2013a). Neural representations not needed - no more pleas, please. *Phenomenology and the Cognitive Sciences*. doi:10.1007/s11097-013-9331-1.

Hutto, D. D., & Myin, E. (2013b). *Radicalizing enactivism: Basic minds without content*. Cambridge: MIT Press.

Kaplan, D. M., & Craver, C. F. (2011). The explanatory force of dynamical and mathematical models in neuroscience: a mechanistic perspective. *Philosophy of Science, 78*(4), 601–627.

Kawato, M. (2008). From "Understanding the brain by creating the brain" towards manipulative neuroscience. *Philosophical Transactions of the Royal Society B, 363*, 2201–2214.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences, 27*, 712–719.

Machamer, P. (2004). Activities and causation: the metaphysics and epistemology of mechanisms. *International Studies in the Philosophy of Science, 18*(1), 27–39.

Machamer, P., Darden, L., & Craver, C. (2000). Thinking about mechanisms. *Philosophy of Science, 67*(1), 1–25.

Nicolelis, M. A. L. (2001). Actions from thoughts. *Nature, 409*, 403–407.

Nicolelis, M. A., & Lebedev, M. A. (2009). Principles of neural ensemble physiology underlying the operation of brain-machine interfaces. *Nature Review Neuroscience, 10*, 530–540.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.
Persson, J. (2010). Activity-based accounts of mechanism and the threat of polygenic effects. *Erkenntnis, 72*(1), 135–149.
Pouget, A., Dayan, P., & Zemel, R. S. (2003). Inference and computation with population codes. *Annual Review of Neuroscience, 26*, 381–410.
Psillos, S. (2004). A glimpse of the secret connexion: harmonizing mechanisms with counterfactuals. *Perspectives on Science, 12*(3), 288–319.
Ramsey, W. M. (2007). *Representation reconsidered*. Cambridge: Cambridge University Press.
Rheinberger, H.J. and Müller-Wille, S. (2010). Gene. *The Stanford Encyclopedia of Philosophy* (Spring 2010 Edition), Edward N. Zalta (ed.), URL = http://plato.stanford.edu/archives/spr2010/entries/gene/
Shea, N. (2012). Reward prediction error signals are meta-representational. *Noûs*. doi:10.1111/j.1468-0068.2012.00863.x.
Shea, N. (2013). Naturalising representational content. *Philosophy Compass, 8*(5), 496–509.
Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction and search* (2nd ed.). Cambridge: M.I.T. Press.
Sprevak, M. (2013). Fictionalism about neural representations. *The Monist, 96*, 539–560.
Tabery, J. (2004). Synthesizing activities and interactions in the concept of a mechanism. *Philosophy of Science, 71*, 1–15.
Waters, K. (2013). Molecular Genetics. *The Stanford Encyclopedia of Philosophy* (Fall 2013 Edition), Edward N. Zalta (ed.), URL = http://plato.stanford.edu/archives/fall2013/entries/molecular-genetics/
Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford: Oxford University Press.
Woodward, J. (2008). Response to Strevens. *Philosophy and Phenomenological Research, 75*, 193–212.
Woodward, J. (2013). Mechanistic explanation: its scope and limits. *Aristotelian Society Supplementary Volume, 87*(1), 39–65.