# Action Guidance is not Enough, Representations need Correspondence too: A Plea for a Two-Factor Theory of Representation

Paweł Gładziejewski

To appear in *New Ideas in Psychology*

**Abstract:** The aim of this article is to critically examine what I call Action-Centric Theories of Representation (ACToRs). I include in this category theories of representation that (1) reject construing representation in terms of a relation that holds between representation itself (the representational vehicle) and what is represented, and instead (2) try to bring the *function* that representations play *for* cognitive systems to the center stage. Roughly speaking, according to proponents of ACToRs, what makes a representation (that is, what is constitutive of it being a representation) is its being functionally involved in preselecting or guiding the actions of cognitive systems. I intend to argue that while definitely valuable, ACToRs are underconstrained and thus not entirely satisfying, since there exist structures that would count as representations according to ACToRs, but which do not play functional roles that could be nontrivially or in an explanatorily valuable way classified as representing something for a cognitive system. I outline a remedy for this theoretical situation by postulating that a fully satisfying theory of representation in cognitive science should have two factors; i.e., it should combine the pragmatic, action-oriented aspect present in ACToRs with an element that emphasizes the importance of the relation holding between a representational vehicle and what is represented.

**Key-words:** representation, interactivism, action guidance, job description challenge, mental models, structural representations, S-representation

## 1. Introduction

The turn of the 20th century in cognitive science will probably be remembered as a time when “embodied”, “enactive”, and “extended” approaches came to play a prominent role in theorizing about, modelling, and studying cognition. Admittedly, there is (still) no universal consensus as to how exactly we should understand these approaches and the way they are interrelated (but see e.g. Goldman, 2012; Shapiro, 2010). However, it seems safe to say that they have two very broad characteristics in common. First, proponents of these approaches usually see themselves as being in opposition to "classical" cognitive science, which construes cognition in terms of rule-based symbolic computation. Second, they criticize the classical view of cognition as too spectatorial or passive, and opt instead for a view that emphasizes that cognition has evolved in order to help embodied agents to control their ongoing interactions with the environments they inhabit.

Among the "orthodox" assumptions of classical cognitive science that are often criticized and discarded by proponents of these new approaches is the idea that cognition involves internal representations. Thus, embodied, enactive, or extended cognitive science seems to be a natural ally of anti-representationalism. Under closer examination, however, it turns out that this diagnosis is an oversimplification. There have been attempts to reconcile representationalism with new approaches in cognitive science (see e.g. Clark & Grush 1999). In this article, I will critically examine a specific strategy for achieving this sort of reconciliation—one that tries to reconceptualize the very nature of representations by postulating that being a representation is constitutively dependent on being somehow involved in guiding the actions of a cognitive system (Anderson & Rosenberg 2004, 2008; Bickhard 1993, 1999, 2004a, 2004b; for an attempt to combine this approach with computationalism in cognitive science, see also Miłkowski, 2013). Throughout this paper, I will call proposals of this sort "Action-Centric Theories of Representation" (ACToRs).

Although I think that ACToRs are, in some important respects, a step in the right direction, I also think that they are fundamentally incomplete. ACToRs are too liberal and underconstrained and thus do not give us a notion of representation that is *explanatorily* nontrivial and valuable. This is because at least some cognitive structures that would have to count as representations according to ACToRs do not meet what William Ramsey (2007) calls the "job description challenge": under closer scrutiny, it turns out that those structures do *not* play functional roles that are recognizably representational, and because of that, they cannot be characterized as representations in an explanatorily illuminating way. Showing that this is the case is my first aim in this article. My second aim is to suggest a way of expanding the notion of representation present in ACToRs so that it can meet Ramsey's challenge. According to my proposal, what we need is a two-factor theory of representation, one that combines the action-oriented or pragmatic element present in ACToRs with the idea that representations *also* owe their representational status to a relation ("correspondence") that holds between the representation itself (the vehicle) and what is represented.

I will proceed as follows. In section 2, I will first describe what I take to be the basic tenets of ACToRs and then take a closer look at two specific theories that are representative examples of the action-centric approach, namely Mark Bickhard's interactivist theory of representation and Michael L. Anderson and Gregg Rosenberg's action guidance theory of representation. In section 3, I will present Ramsey's idea of the job description challenge. In section 4, I will try to show that the notion of representation contained in ACToRs is too liberal and underconstrained to meet the job description challenge. In section 5, I will suggest that this problem can be dealt with by extending ACToRs to create a two-factor theory of representation. I will also present a very sketchy outline of how this sort of two-factor theory might (and should) look.

## 2. Action-Centric Theories of Representations

### 2.1. ACToRs: core ideas

It might be useful to introduce ACToRs by pointing to what they are *opposed* to. Proponents of the action-centric approach often claim that their proposals are based on the rejection of a certain way of thinking about the nature of representation, one that is deeply embedded in today's mainstream philosophy and cognitive science. As Mark Bickhard (2004b) puts it, this way of thinking construes representations as *encodings* or *correspondences*, and it can be expressed using the metaphor of "impressions" left by a signet ring (the world or what is represented) in a piece of wax (the representation). Correspondence-based theories see representations as codes whose constituents are mapped on to constituents of the represented domain. Correspondences are supposed to be established by some sort of (natural) relation that holds between the representation and what is represented. But what sort of relation is this? This is a broad subject, but suffice it to say that causal-nomological dependence, asymmetric causal dependence, or isomorphism are some of the candidates that have been proposed in the contemporary literature. Michael L. Anderson and Gregg Rosenberg express a similar diagnosis to Bickhard when they claim that the problem with many contemporary theories of representation lies in the fact that they are input-focused, meaning that "they give too much importance to the ways in which the environment affects the organism to endow its states with representational meaning" (2008, p. 56). To put it very broadly then, ACToRs are opposed to a very general idea about the fundamental nature of representation, namely the idea that what is constitutive of being a representation is a correspondence between the representation itself and what is represented.

Characteristic of proponents of ACToRs is that, instead of proposing yet another correspondence-establishing relation, they attempt to make something of a paradigm shift in our thinking about what representations really are. If most "classical" theories are indeed input-focused, then ACToRs can be described as trying bring the representation's *output*—i.e., the relationship between representation and its *user*—to the center stage. From this point of view, of crucial importance for our thinking about the nature representations is the fact that representations are *for* their users, with all their practical purposes. To put it more precisely, we could say that proponents of ACToRs approach the subject matter in the following way. First, they ask what it is that representations *do* for their users, or what is the "business" of using representations. Second, they treat an answer to this question as a basis for their positive theory of representation, in accordance with Anderson and Rosenberg's claim that "representations *are* what representation *do*" (2008, p. 56, emphasis added).

So what function do representations serve for their users? According to ACToRs, their role consists in controlling or guiding the user's actions. Thanks to representations, cognitive systems have the ability to practically orient themselves in the world, perform actions that are adaptive given the circumstances, or (pre)select one action among many that are potentially available at a given moment.

One very important clarification about what I take ACToRs to be committed to is in order. ACToRs can be interpreted in two very different ways. According to a *weak* interpretation, ACToRs simply give an account of the function that representations play for representation-using systems. According to a *strong* interpretation, (1) ACToRs are theories of what *constitutes* representations—what makes them representations—and (2) they put forward the thesis that representations are constituted by their function in providing guidance for action. The weak and strong interpretations are clearly very different. Imagine an analogical situation, in which someone claims that the function of cars is to enable people to cover long distances. On one hand, this is a trivial truth about what cars are for. On the other hand, if you interpret this claim strongly—i.e., as saying that what *constitutes* a car is playing a role in enabling people to cover long distances—then you end up with a thesis that is obviously false, since there exist other artifacts that carry out the same function. Now, my aim here is not to challenge an idea that I take to be fairly unproblematic, namely that representations serve to guide or preselect actions. I will not argue against a weak interpretation of ACToRs. I also do not think that the weak interpretation is what proponents of ACToRs argue for. What they mean to defend, I think, is the strong interpretation of their thesis. And it is this version of action-centrism with which I take issue. What I want to bring into question here, then, is the idea that what makes something a representation is the fact that it is involved in controlling the actions of cognitive systems.

The discussion so far has given a very cursory introduction to ACToRs by pointing out the basic ideas that they have in common. Now I want to take a closer look at two theories that I take to be emblematic of the action-centric approach: Bickhard's interactivist theory and Anderson and Rosenberg's action guidance theory. I will present both theories in subsections 2.2 and 2.3, respectively. However, it needs to be stressed at the outset that both theories are conceptually complex in a way to which I cannot fully give justice here. For example, Bickhard presents his proposal against a broader conceptual and theoretical background of process ontology and an "interactivist" view of cognition as such (see Bickhard, 2009). In addition, both Bickhard (1993, 1999, 2004a, 2004b), as well as Anderson and Rosenberg (2004, 2008) put a lot of effort into developing their respective theories in a way that avoids presupposing representational notions when explaining representation. In my reconstruction, I will pass over many of these details and instead concentrate on what I take to be the essential theses of both theories. I will also charitably assume that both theories avoid vicious circles or metaphysical assumptions that are in any way problematic.

### 2.2. Bickhard's interactivist theory of representation

Bickhard builds his theory on the idea that representations owe their normativity—the fact that they are assessable for truth or accuracy—to the normativity of (inter)actions (the following reconstruction is based on Bickhard, 1993, 1999, 2004a, 2004b). Actions have certain conditions of success, and the normativity of representations is derivative, explanatorily and metaphysically, from that practical normativity. But what does the normativity of actions consist in? Bickhard's answer to this question points to the fact that all living systems need to maintain themselves by staying in a state that is far from thermodynamic equilibrium. Now, contrary to some self-maintaining systems like a candle flame—which keeps itself in existence by vaporizing wax into flammable gases, but is able to do so only under one set of specific external conditions—*cognitive* systems have the capacity for *recursive* self-maintenance. That is, the latter can maintain themselves over a range of changes in environmental conditions. Cognitive systems can differentiate environmental circumstances and change their course of action accordingly, keeping themselves in a far-from-thermodynamic equilibrium state despite changes in environment. For every action (type), there are (types of) conditions under which this action contributes to the organism's maintenance. These are, for Bickhard, conditions of this action's success.[1]

Bickhard observes that in cognitive systems of appropriate complexity, it will usually be the case that (1) there is more than one possible action to perform in given circumstances, and (2) the system thus faces the challenge of preselecting one of those actions in a way that maximizes its self-maintenance. In such a situation, organisms will not simply trigger a given action in given circumstances, but will rather develop an ability to set up what Bickhard calls *indications of interactive potentialities*. Without delving into the details of Bickhard's definition of indication, we may roughly say that indication I of some potential action A in circumstances C is an internal state (or process, if we choose to stay closer to Bickhard's metaphysical framework) which is produced as a result of the occurrence of C and which enables the system, or makes it possible for the system, to perform action A (for details, see Bickhard, 1993). Usually more than one action will be available in a given situation, so that more than one possibility for action will be indicated[2].

According to Bickhard, indications of interactive potentialities are a fundamental or basic form of representation. It needs to be stressed, though, that what makes them representations is not how they are related to the external world, but rather the role they play with regards to *(inter)actions*. They owe their status as representations to how they enable the organism to act in an adaptive, anticipatory way. Every indicated action has conditions of success associated with it, that is, conditions that *should* occur if this action is to contribute to the system's self-maintenance. As Bickhard puts it, indicated (inter)actions "dynamically presuppose" that conditions occur which are needed for those actions to be successful. It is those dynamic presuppositions that constitute or determine representational *content*. These

[1] For example, a bacterium may have the ability to perform two actions: swimming and tumbling (Bickhard 2004a). Swimming enables self-maintenance when the bacterium is moving up a sugar gradient, and so this is this action's condition of success. Tumbling contributes to self-maintenance when the bacterium finds itself moving down a sugar gradient, and so this is this action's condition of success.

[2] Sometimes Bickhard (e.g. 1999) also claims that, in addition to indications of interactions themselves, the interaction's internal *outcomes* need to be indicated as well. If this claim is added to the theory, then what is needed to set up an indication of interactive potentiality is not only an indication of a possible action itself, but also an indication of internal states that the former is predicted to produce.

contents are not explicitly encoded for Bickhard, but rather are implicit in the dynamical presuppositions of indicated actions.

Two important additional remarks are in order. First, Bickhard claims that the abovementioned story only explains the most fundamental or rudimentary form of representation, namely interactive representation. As I understand the theory, these are representations with simple contents like "Circumstances affording action X, now". However, Bickhard also claims (and attempts to show, see esp. Bickhard, 2004b; Bickhard & Terveen, 1995) that other, more sophisticated forms of representation—like representations of stable physical objects or abstract entities—are built upon the basis of interactive representation.

Second, Bickhard claims that his interactivist theory has one very important advantage over any correspondence-based theory. This is that only the former can explain not only how representational error is possible, but also how s*ystem-detectable representational error* is possible (see especially Bickhard, 1999). For Bickhard, explaining how it is possible for a cognitive system to detect the fact that it is using a false or inaccurate representation is an important criterion of adequacy for any theory of representation. This is also where he thinks all theories based on an idea of representation as correspondence uniformly fail. Animals seem to have no way of telling whether a given correspondence-establishing relation (causal co-variation, asymmetric dependency, isomorphism, etc.) holds between the representation they are using and what they are representing. Furthermore, from the point of view of correspondence-based theories, error detection would need to generate a regress: to detect an error of representation $R_1$, the system would need to form an independent representation $R_2$ of whether the correspondence-establishing relation holds between $R_1$ and the world; but now the problem reappears, since similar sort of story would be needed to explain the ability to detect an error in $R_2$ (which would call for yet another representation $R_3$); and so on. From an interactivist point of view, explaining error-detection seems to be achievable in a rather straightforward way. The representation is false if its dynamic presuppositions are false. A representation-user can detect this falsehood by detecting that an *action* based on this representation has failed. Representational error is detectable through action failure. So whenever an action that was based on a representation fails, detecting the occurrence of *this* fact is a way of telling that the representation was false, which may in turn be potentially useful for error-based learning.

### 2.3. Anderson and Rosenberg's action guidance theory of representation

I will now turn to Anderson and Rosenberg's (henceforth A&R) action guidance theory (the following reconstruction is based upon Anderson & Rosenberg, 2004, 2008). At the heart of A&R's proposal is the idea that the function of representation is to guide the actions of representation-users. Representations are entities whose job it is to enable natural and artificial cognitive systems to perform intelligent, adaptive actions with respect to their environments. When presenting their case, A&R are mostly concerned with unpacking this very broad and intuitive idea in a technical and detailed manner. So, according to their technical proposal, "a mental token T (of some given type) represents an entity E for subject S, just in case tokens of that type are standardly used by S to guide its actions with respect to E" (Anderson & Rosenberg, 2008, p. 67). Let us take closer look at this characterization.

For A&R, a T token guides the actions of S by "making its features available to the subject's motor systems and rational control processes for use in making discriminating choices between possible actions or possible ways of executing actions" (Anderson & Rosenberg, 2008, p. 68). Importantly, the category of representation-guided actions is supposed to include *both* motor actions and cognitive actions. By saying that T tokens are "standardly used" by S to guide its actions, A&R mean that S uses T tokens to guide its actions with respect to E in virtue of having an "enduring conscious preference or conditioned reflex" to use T tokens to guide its actions. A&R also put a lot of effort into characterizing technically what it means for an action to be performed "with respect to" some entity E. To cut a long and complex story short and make it (relatively) simple, three conditions need to be met for an action to be performed with respect to E (the following discussion covers motor actions only, although A&R treat cognitive actions in a very similar manner). First, the action should be susceptible to rational interpretation and have a "motivational reason" behind it. Second, the motivating reason behind this action should be an "assumption of information" about E, i.e. the action should be such that it unfolds as though it were based on assumption that T tokens carry information about E (the system in question behaves "as if" it assumed that T tokens encode information about E). Third, E should be the focus of the expected change that this action is to bring about; more precisely, E should be the entity that is continuously monitored by feedback channels as the action is performed.

Action guidance theory can be illustrated with the following simplistic example. Consider a very simple cognitive system S, whose whole life boils down to approaching some entities (say, food and reproductive partners) and avoiding others (say, predators). S contains a very simple mechanism, composed of two components: (1) M, which sends "approach" or "avoid" motor commands to the effectors, and (2) D, which is in the business of affecting M in such a way that it actually sends one of those two commands to the effectors, depending on whether D itself is in a state $D_1$ (which activates approaching) or $D_2$ (which activates avoidance). In other words, D "decides" what should be done and M moves the system in one of two ways. The whole process is set up as though S acted on an assumption that D encodes information about external states of affairs. Suppose that on a given occasion S stumbles upon a token of the type O (say, a piece of food), D goes into state $D_1$ and thus causes (through affecting M) S to move in the direction of O. Suppose further that D in the state $D_1$ is *standardly* used to make S approach Os. If this is the case, then D in state $D_1$ could be described as guiding S's actions with respect to Os, and thus as *representing* Os for S. In other words, $D_1$ would then represent simple content like "Food here, now".

One last observation to make about action guidance theory is that A&R follow Bickhard when they claim that an important strength of their theory is its ability to explain system-detectable representational error. The way A&R handle this subject is basically the same as Bickhard. Representational error consists in the fact that an action that was guided by a given representation "failed in its intent, and it failed partly or wholly because of the guidance provided by that representation" (Anderson & Rosenberg, 2008, p. 78). This error in representation can be detected when an action guided by this representation fails.

## 3. Ramsey's job description challenge

As was stated at the outset, my aim in this article is to challenge ACToRs. It is important, though, to be explicit about *how* I intend to do this. I will approach my goal from a specific angle, one that is different from the way theories of representation are usually treated in the literature—namely as attempts at naturalizing intentional *content*. Instead of asking whether ACToRs give us a good theory of how intentional content is determined, I want to concentrate on their *explanatory* value for cognitive science. More specifically, my intention is to ask whether representations postulated by ACToRs actually play their explanatory roles *qua representations*; or whether explanations of cognitive phenomena that posit internal structures meeting ACToRs' criteria for being a representation actually earn their status as genuinely *representational* explanations.

My discussion will draw on a strategy of evaluating the explanatory status of representational posits in cognitive science that was introduced by William Ramsey in his book *Representation Reconsidered* (2007). Ramsey's proposal stems from a suspicion that the term "representation" is often used by cognitive scientists in an excessively liberal and unconstrained way, which deprives it of any true explanatory value. When explaining phenomena of interest, cognitive scientists show a tendency to call all sorts of internal structures "representations", including structures for which it is far from clear why they should be classified that way at all. The problem is *not* that those structures do not play an important role in cognitive-scientific explanations, but rather that they do not play their explanatory roles *as representations*.

Ramsey (2007) postulates that we need a principled way of judging the real explanatory value of different notions of representation routinely used in cognitive science. His positive proposal is founded on the idea of the job description challenge (henceforth JDC). This idea is based on an assumption that representational explanations in cognitive science consist in describing cognitive mechanisms whose components are functionally involved in representing something. That is, something explains a given phenomenon as a representation because it functions as a representation in a mechanism responsible for that phenomenon (this assumption is in line with the mechanistic outlook on the nature of cognitive-scientific explanation, see e.g. Bechtel, 2008). Meeting the JDC requires showing how exactly a given internal structure or state posited as a representation actually serves or functions *as a representation* in a given cognitive mechanism.

The procedure of using JDC runs roughly as follows (see Ramsey, 2007). One starts by examining the conceptual and explanatory practice of cognitive scientists in order extract a specific notion or a way of understanding representation that is implicitly or explicitly present in this practice. In doing so, one must concentrate on the functional roles that are attributed to purported representations, that is, their "job description". Second, one needs to take a closer look at this job description and ask whether states or structures that meet this job description can be classified—in an intuitive, natural and understandable manner—as playing the role of representation. Does it make sense to describe these structures as *standing-in* for something? Or does the intentional content we may attribute to them determine their functional role within the system or mechanism in which they are embedded? According to Ramsey, if answers to questions like these are positive for a given notion of representation, then this notion meets the JDC. Using this notion should give some real explanatory purchase or enable us to understand phenomena in way that is impossible without it. On the other hand, if the answers are negative in a given case, then we are dealing with a notion of representation that fails to meet the JDC. This notion turns out to be superfluous and devoid of real explanatory value. We could just as well do without it in our explanations and not lose theoretical insight into the phenomena of interest.

Before I return to discussing ACToRs, let me illustrate the procedure outlined above with a concrete example taken from Ramsey's book (this example will also prove useful further down the line in my argumentation against ACToRs). Among a number of cognitive-scientific representational notions that Ramsey (2007) examines there is what he dubs the "receptor notion". Ramsey thinks this is one of the most prevalent ways of conceptualizing representation in cognitive science (particularly in connectionist modelling and cognitive neuroscience). Receptors are supposed to represent in virtue of the fact that they reliably co-vary with what is represented. For example, the role of *representing* a stimulus S is often ascribed to an individual neuron in the visual cortex of an animal on the basis that its activity *reliably co-varies* with presence of S in the animal's visual field.

Do receptors—or, rather, our concept of representations as receptors—meet the JDC? Ramsey's discussion of this issue is quite complex and detailed, so for the sake of space, I will only present the gist of his analysis (for details, see Ramsey, 2007, pp. 118–150). The first thing to note is that the basic notion of representations as receptors is far too liberal. There are numerous reliable co-variances in the universe that, as it seems, do not represent anything or do not play any representational function. Thus, if we were to accept receptors as genuine representations, we would be faced with pan-representationalism, which would in effect trivialize representational explanations. One plausible way of avoiding this problem is to add a teleological element to the receptor notion. Following Dretske (1988), we might characterize receptors as structures that (1) reliably co-vary with some state of affairs and (2) are functional for some larger system or mechanism because of this co-variance. For example, a neuron in a visual cortex may not only co-vary with the presence of a stimulus, but also be functional (say, by producing appropriate behavior, like avoidance) for the organism precisely because it co-varies with it. However, even now the receptor notion is still too liberal, since we can point out counterexamples, that is, structures in the world that meet both those criteria but clearly do *not* function as representations. To take one of numerous counterexamples given by Ramsey, the firing pin's *function* is to activate the discharge of the round by being reliably activated by pulling the trigger, but the firing pin obviously does not *represent* that the trigger was pulled.

It seems, Ramsey observes, that there is a one last way to rehabilitate the receptor notion, namely by adding the concept of *information* to the picture. It might be said that a receptor is a representation when, in virtue of co-varying with a state of affairs, it encodes information about that state of affairs and is functional for a larger system because of the information it encodes. Ramsey goes to great lengths to neutralize this move as well. Roughly, he argues that in order to defend her case, the proponent of the receptor notion should be able to characterize the relationship between co-variance and information in such a way that it is possible to show that the *latter*, and *not* the former is what makes the receptor functional (otherwise we fall back on the firing pin counterexample). But Ramsey argues that this is *only* possible if we assume that the information is used by a full-blown (human-level) intentional agent who is able to understand the entailment relation holding between the receptor and the (represented) state of affairs. However, if this is so, then we cannot make use of the receptor notion to understand representations in the sense that is of interest to cognitive scientists—that is, *mental* representations which are supposed to be located *inside* cognitive systems and used *inside* them—since we cannot (under the threat of homuncular fallacy or vicious regress) postulate agents with human-like interpretative abilities populating the insides of organisms.

The upshot of Ramsey's discussion of the receptor notion is this. Purported receptor representations are not in fact in the business of representing anything. The way they function is more similar to the way gear-wheels function than to the way the things that we pre-theoretically recognize as representations—like maps, fuel gauges, or sentences of natural language—function. For example, there is no sense in which receptors stand in for anything. No doubt receptors causally mediate between states of affairs, but causal mediation is obviously not sufficient for representation. Thus, according to Ramsey, the receptor notion does not meet the JDC.

## 4. Do ACToRs meet the JDC?

### 4.1. ACToRs and the JDC

Notice the affinity between Ramsey's project in *Representation Reconsidered* and the aim of proponents of ACToRs. Both projects explicitly concentrate on the functional role that representations (purportedly) play within a cognitive system. In other words, when developing their theories, proponents of ACToRs are asking precisely the type of question we should be asking, according to Ramsey, if we want an explanatorily valuable notion of representation, namely: What is it that representations *do* something *for* their users? Because they explicitly characterize representations in functional terms, it seems that ACToRs can be quite straightforwardly confronted with the JDC. All we need ask is whether internal structures that meet the "action-centric" functional criteria for being a representation in fact serve roles that are recognizably and in some illuminating or nontrivial way representational.

While I think that they are a step in a right direction when it comes to providing us with an explanatorily valuable notion of representation, as they stand, ACToRs fall quite short of meeting the JDC. In subsections 4.2 and 4.3, I will show why I think this is the case by discussing the shortcomings of, respectively, Bickhard's interactivism and A&R's action-guidance theory. In subsection 4.4, I will consider a possible answer to my worries and show why this answer is unsatisfying.

### 4.2. Bickhard's interactivist theory: a counterargument and a counterexample

Let me start with Bickhard's interactive theory. I think there are two arguments for the claim that the notion of representation embedded in this theory does not meet the JDC. One of these is purely conceptual, while the other is based on an counterexample drawn from actual empirical work done in cognitive science.

The first argument rests on the observation that there is no non-question-begging way of showing how exactly the function played in a cognitive system by indications of interactive potentialities can be qualified as representational. According to interactivism, those indications are internal structures or processes that (1) are activated or arise before the organism engages in particular action (say, predator avoidance); (2) can, and sometimes will, lead to the organism actually performing this action; (3) have semantic or representational contents that are determined by the conditions of success (dynamic presuppositions) of the action to which they can lead. Note that according to the theory, the representational status of indications of interactive potentialities is fully constituted by the role they play with respect to guiding or preselecting action. In other words, indications have property (3) in virtue of having properties (1) and (2). We may say, then, that the whole of Bickhard's project is to explain representational normativity with the normativity of action. However, we need to keep separate two alternative ways in which this could be done. According to one possible strategy, we might say that the normativity of action is somehow explanatorily and/or metaphysically relevant to the normativity of representation. This way of framing the issue invites a theory according to which something is a representation in virtue of being appropriately related to possible actions. But this does not mean that the appropriate connection to a possible action is, by itself, (necessary and) sufficient to make something a representation. My intention is not to argue against this idea. In fact, I take it to be on the right track (see section 5).

According to the second possible strategy of explaining representational normativity in terms of the normativity of action—we may call it the "*reductive*" strategy—the former is, as the saying goes, "nothing over and above" the latter. From this perspective, an action having conditions of success is the same as a representation (indication of interactive possibility) having content. However, this way we do not account for the nature of representations, but rather render representations explanatorily useless. Everything said using representational talk could be more economically said by reference to actual or possible (inter)actions and their conditions of success. There is no rationale for introducing representations into the picture. The simple fact that some internal activity could eventually cause an action whose success depends on environmental conditions gives, by itself, no leverage to the idea that this activity *represents* those conditions. Now, I think that by simply identifying dynamic presuppositions with (implicit) contents, Bickhard is committed to this second, reductive strategy. Given that this strategy is unsuccessful, I see no reason for claiming that interactivism gives us an explanatorily valuable notion of representation.

A proponent of the interactivist theory of representation might reply to this by pointing to the fact that interactive representations are *not* explicit, but rather represent their contents only *implicitly*. By giving this sort of response, the proponent of interactivism seems to embrace what Ramsey calls a "tacit" notion of representation (see Ramsey, 2007, pp. 151–187). Supposed tacit representations are not encoded in some specific neural or computational structures, but are rather non-locally "embodied" in the internal functional architecture of a cognitive system as a whole (e.g. in the pattern of connection weights between the nodes of a neural network). Although it is not possible to delve into this topic right now, it needs to be stressed that it is highly debatable whether the notion of tacit or implicit representation meets the JDC. In fact, Ramsey (2007) presents a strong case for claiming that there is no way in which tacit or implicit "representations" are in fact in the business of representing anything. He argues that attributing tacit representations to a given systems boils down to saying that the system as a whole has some dispositional profile (e.g. it reacts differently to different categories of objects). But this does not entail that there is anything even remotely functionally resembling a representation *inside* the system.

Now, I do not want to simply preclude the possibility that there are such things as tacit or implicit representations. However, I think that even if someone categorizes something as an *implicit* representation, she needs to provide a good rationale for using representational talk at all. Does Bickhard's theory give us such a rationale? If so, what might it be? I think that the interactivist answer to these questions would probably run as follows. Each indicated action has some dynamic presuppositions, i.e. conditions of success. Those dynamic presuppositions are, according to Bickhard's theory, what determines the representational content. However, there is nothing in the system that would explicitly represent this content. So the environmental conditions that constitute the dynamic presuppositions of indications of interactive potentialities must be represented only implicitly. Notice, however, that this sort of answer manifestly begs the question. It simply gives us no good reason to postulate representations in the first place. It merely presupposes representations. Indications of interactive potentialities enable the organism to perform actions that have conditions of success. This is the *only* way in which those indications are related to external conditions. Saying that they *also represent contents*, even implicitly, simply adds nothing genuinely new or explanatorily valuable to the picture. Thus, again, there is no theoretical or explanatory benefit in postulating representations here.

Let me now turn to the second, more empirically-oriented argument against interactivism. The point is simple: there exist clearly non-representational explanations in cognitive science that nonetheless *do* meet the criteria for being representational put forward by interactivism. Explanations of this kind constitute counterexamples for interactivism and show that the notion of representation embedded in Bikchard's theory is too liberal to meet the JDC.

To see this, consider a famous work by Randall Beer (2003) that is often cited as a model example of anti-representationalism in cognitive science. In this study, Beer describes a virtual organism—a result of many generations of virtual natural selection—inhabiting a two-dimensional virtual environment. The agent's movements are controlled by a three-layer neural network attached to a virtual eye. The organism is able to perform horizontal movements in order to engage in one of two possible actions: catching or avoidance. Its whole virtual "life" consists of facing the challenge of avoiding diamond-shaped objects and catching circles, with both kinds of objects falling on it from above. Importantly, the agent stumbles upon not only diamonds and circles, but also hybrids of these two shapes whose categorical membership is more vague.

In his article, Beer (2003) has shown that although the agent exhibits behavior that should earn it the status of a minimally *cognitive* system (it had an ability for active categorical perception of objects falling on it), the network that controls its movement does *not* make use of any internal structures that would deserve the label "representations". Beer's reasoning for this latter claim is familiar: his dynamic-systems-based analysis of the mechanism controlling the agent's actions showed that there is nothing inside this mechanism that could be nontrivially described as performing a representational *function*, thus rendering representations explanatorily useless in this case. The virtual agent does not make use of any kind of model of its environment. It does not make use of any kind of diamond- or circle-detectors either, as there are no internal states that would systematically and exclusively co-vary with either of those shapes. Although the system's behavior as a whole could be usefully described using intentional-sounding categories—like "deciding" or "intending" to avoid objects—there are no internal, causally relevant structures governing its actions that correspond to those categories. All in all, to put it in Ramsey's terms, Beer's virtual agent is a non-representational system because there is nothing inside it that would meet the JDC (i.e. play a truly representational function).

Notice, however, that despite being a non-representational system, Beer's agent nonetheless seems to meet interactivism's criteria for being representational. First, it has the ability to select one of two possible actions, each of which has some conditions of success (dynamical presuppositions) and which could potentially fail, as e.g. when the system "decides" to catch a diamond. Second, the agent is controlled by a neural network that makes use of internal states and processes that *meet* the criteria for indicating interactive potentialities. Whenever the organism scans its environment, there are ongoing patterns of activity in the middle layer of virtual neurons (those that send motor commands to the third, motor layer) that could and sometimes eventually would cause the agent to perform one of the actions available to it. The patterns leading to each action seem to "compete" with each other for control (as Beer [2003, p. 221] puts it, his results "[...] suggest that the 'decision' [about how to act] is repeatedly made and unmade as the agent and the object interact" until the organism eventually "commits" and actually performs one of the actions. It seems, then, that interactivism would have us think that in this case there are patterns of internal activity that (1) precede action and (2) could potentially lead to the agent performing a particular action; thus we should say that these patterns indicate interactive potentialities that therefore (3) *represent* the action's conditions of success. It follows that Beer's virtual agent is a non-representational system that is categorized as representational by Bickhard's theory. This shows that the notion of representation present in interactivism is too liberal: it lets *too many things* count as representations. Of course, someone might claim that interactivism shows that our evaluation of the example should be reconsidered and that we are dealing with a representation-using system after all. But this move does not seem warranted. It would require stretching the notion of representation beyond the point of usefulness. As Beer (2003, p. 239) himself puts it, if the mechanisms of the sort he describes "look nothing like representations, and they act nothing like re-presentations, then they are not representations, and continuing to call them representations is just going to confuse everyone".

### 4.3. Anderson and Rosenberg's action guidance theory and the frog's fly detection mechanism

Let me now turn to A&R's action guidance theory. I think this theory is also open to counterexamples. However, this time instead of looking for a counterexample in cognitive science, I would like to critically re-examine the case that A&R themselves present as an *illustration* of their theory, namely the famous fly-catching mechanism in frogs (Anderson & Rosenberg, 2008; see also Lettvin et al. 1959). Whenever a fly or a fly-like moving object is placed within a frog's field of vision, retinal ganglion cells are activated, which in turn activate neurons in the frog's optic tectum. Neurons in the optic tectum are connected with neural structures that control the frog's behavior. The mechanism is set up such that it causes the frog to orient its head in the fly's (or fly-like object's) direction and snap its tongue to catch it. According to A&R (2008), this is a simple representational mechanism. What justifies interpreting it in representational terms is that the activity in the optic tectum is functionally involved in guiding the frog's actions in a way that meets A&R's functional criteria of being a representation.

Why do I think that the frog example can in fact be used to show that action guidance theory does not meet the JDC? Basically, the idea is that by treating the fly-detection mechanism as representational, A&R's theory is faced with the same problems that bug the receptor notion of representation discussed in section 3. The fly-detection mechanism employs an internal structure that is reliably activated by the presence of flies (or fly-like objects) and whose function it is to initiate actions with respect to that which activated it. In other words, it acts a sort of action-guiding *receptor* (see Ramsey 2007). It turns out then that receptors—or, rather, those receptors that are involved in controlling the actions of cognitive agents—constitute at least a subset of structures that meet the functional criteria of being a representation put forward by A&R in their action guidance theory. However, for reasons I have presented in section 3, it seems that receptors do *not* play a *representational* function in cognitive systems. We may simply conceptualize them as internal causal mediators without losing any explanatory power or insight into the workings of a given system. The same applies to frog's fly-detection mechanism. But this means that the neurons in frog's optic tectum are an example of a nonrepresentational structures that nonetheless meet A&R's criteria. If so, then it follows that action guidance theory would have us treat as representations structures that do not really perform a function of representing anything. Action guidance theory is too liberal; the way it construes representations falls short of meeting the JDC[3].

When presenting their theory, A&R (2008) seem to anticipate this sort of criticism and attempt to answer it preemptively. They claim that we do have a reason to think that fly-detection in frogs in achieved by employing representations, and not simply by employing a multi-step causal chain. To show this, they contrast (what they take to be) representation-guided actions of frogs with a nonrepresentational, purely casual action guidance in the case of slime-mold's phototaxis. In short, they claim that what distinguishes the former from the latter is that in the frog's case there is a *potential decoupling* of stimulus and response.[4] There

[3] Interestingly, the fly example is occasionaly also used by Bickhard and his co-workers to illustrate the interactivist approach to representation (see e.g. Bickhard & Campbell, 1996, Bickhard & Terveen, 1995). On this construal, because the activation of neurons in the optic tectum enables the frog to perform tongue flicking action (i.e. it increases the probability that the frog will in fact perform this action), this activity indicates a certain interactive potentiality, and thus constitutes a representation. However, conceptualizing this example in interactivist terms does not neutralize Ramsey's critique any better than conceptualizing it using action guidance theory. For example, the fly catching mechanism remains a non-representational causal mediation mechanism even if we accept that, as the proponents of interactivism point out, the neural activity in the frog's optic tectum causes tongue-flicking action only *ceteris paribus* (for example, only if the frog is not at the same time exposed to a shadow of a predatory bird; see Bickhard & Campbell, 1996). I have decided to discuss the fly catching specifically in context of A&R's theory because of how these authors develop this example by introducing the notion of potential decoupling (see main text for details).

[4] In fact, there is one more difference between these two cases that A&R mention: whereas in the slime mold case the environment directly controls behavior, in the frog's case the stimulus first *registers* inside the brain, and then action is controlled by this internal registration (rather than directly by the stimulus). One of the reviewers of this article pointed out that it might be argued that, by assuming that registration is involved, A&R are *not* in fact advocating a purely action-centric theory, but rather advocate a *two-factor* theory of sorts. After

are two stages of processing at which this decupling can occur. First, internal registration—defined as "a distinct and characteristic inner state, typically formed in response to a certain kind of bodily (sensory) change, and taken up into a behavioral control system" (Anderson & Rosenberg, 2008, p. 62, n. 8)—can be decoupled from external stimuli. For example, in the frog's case, neurons in the optic tectum can register a whole lot of stimuli—like bits of paper or dots on screen—other than just flies. Because these stimuli may have nothing in common physically, we cannot, according to A&R, explain the frog's behavior simply by reference to the *causal features* of the stimuli. Second, there can be a decoupling between the registration and the behavioral reaction it causes. For example, if we unilaterally remove the frog's optic tectum, eventually its optic tract will innervate the intact part of the optic tectum. As a result, the pattern of the frog's reactions to stimuli will be inverted and the frog will snap its tongue at the spot where the mirror image of a fly would be.

I do not think that potential decoupling is enough to delineate representational from nonrepresentational action-guiding mechanisms. First, the existence of decoupling between stimulus and registration by no means entails that we need *semantic* notions to explain a system's reactivity to stimuli. For example, although there is a range of objects that can potentially be registered by a frog's optic tectum, there is no reason *not* to think that these

all, by employing the frog example, they point not only to how the (postulated) representations in the optic tectum guide action, but also to the fact they systematically co-vary with the presence of flies, which may be considered as establishing some kind of correspondence between the representational vehicle and what it represents. Three things need to be said in this context. First, despite the fact that this particular example includes co-variation between the representation and the represented object as part of the story, calling A&R's theory "two-factor" would be unwarranted. These authors do not claim that any sort of correspondence is in any way *necessary* to establish a representation. On their view, what is constitutive of representation is solely how the vehicle guides action with respect to what is represented. Nonetheless, this does not preclude the vehicle (sometimes) being *in addition* causally or co-variationally related to what is represented. So detectors still fall within the category of representations according to A&R. Thus, showing that these structures are in fact not representational still constitutes a countexample to A&R's theory. I have decided to discuss the frog example simply because A&R themselves employ it to illustrate their version of an ACToR. Second, I think the criticism presented here remains perfectly conclusive even when we stick to A&R's theory and completely ignore the fact that the neurons in frog's optic tectum co-vary with the presence of flies. If anything, ignoring this fact would make the case for frogs using representations to detect flies even *weaker*. Third, as I see it, adding registration to the story makes a difference in quantity, not quality, and it boils down to the fact that the causal chain leading from the environment to behavior is longer in the frog's case than in the case of, say, the slime mold. This is not enough to render the frog's fly-detection mechanism representational (see a related discussion in Ramsey, 2007, pp. 189–203).

objects *do* share some *causal*, or causally relevant, properties; namely, properties like being (appropriately) small, being dark, or exhibiting a certain movement patterns. All objects that activate a frog's snapping behavior have these properties in common and activate it in virtue of having these properties.

Second, the decoupling of registration and response can be easily understood as an abnormal reversal or reshuffling of a biological causal chain. By removing part of a frog's causal optic tectum, an optic tract connects with the part of the optic tectum with which it would not be connected normally. Part A of a causal chain, which would normally produce effect B, now produces effect C. The same sort of reversal of a causal chain can be rather trivially achieved in non-biological artifacts like light switches or taps, but nobody would suspect these artifacts of being representational because of this.

Third, even if both types of decoupling co-occur in a given mechanism, this still gives no leverage to a representational, as opposed to purely causal, explanation of this mechanism. Take the frog example again. The mechanism that subserves the frog's snapping behavior can be quite easily explained purely associatively, as a sort of *reflex*. In the frog's natural environment, the appearance of food is systematically correlated with the appearance of stimulus that has a certain color, size and exhibits certain movement pattern. The frog has a reflex mechanism that causes it to automatically snap its tongue in the direction of the stimulus. This reflex can be activated by objects other than food, as long as they have the appropriate features. Whenever this happens, the frog's *action* is *unsuccessful* because it fails to provide the frog with food. Furthermore, if we damage this mechanism in a certain way, it can partially regenerate, but (because of such and such causal or physiological factors) the resulting pattern of behavioral reactions to stimuli may become reversed in interesting ways. I think that this sort of explanation tells us—though admittedly in a very general way—all we need to know in order to gain a basic understanding of a frog's snapping behavior. Note that it does not invoke representations or semantic properties. Instead, it exclusively refers to a reflex-based causal chain and a purely practical (nonrepresentational) connection between the frog's actions and environmental conditions.

Before I move on, let me be clear about exactly what I am arguing for. I do not doubt that potential decoupling distinguishes some action-guiding internal structures from others. Furthermore, I do not want to put into question the idea that this distinction can be theoretically or explanatorily interesting and useful. As A&R (2008) show, a control mechanism that is potentially decoupled can be more flexible than one that is not, and therefore the former can have some adaptive advantages over the latter. Nonetheless, I think it is far from clear how potential decoupling could distinguish *representational* from *nonrepresentational* action-guiding structures. In fact, I think that the frog example shows that these two distinctions (potentially-decoupled/not-potentially-decoupled vs. representational/nonrepresentational) are orthogonal.

### 4.4. Error detection to the rescue?

I now want to investigate one potential line of defense that a proponent of action-centric approach might use against my argumentation. According to this defensive strategy, the whole discussion of this section is inconclusive because I have failed to take into account one crucially important element of ACToRs. As mentioned in subsections 2.2 and 2.3, Bickhard and A&R claim that the strength of their respective theories partially lies in the fact that they can account for the ability to detect representational error. According to ACToRs, representational error is recognized or detected on the basis of failure of action. Now, one might propose that this ability to explain error detection not only gives ACToRs an advantage over correspondence-centric theories of representation, it also enables them to meet the JDC. According to this line of reasoning, what gives some internal state or structure a representational status in not only the fact that it is appropriately involved in guiding or pre-selecting actions, but *also* the fact that when it fails in performing its action-related function, it fails in a way that is detectable *for* a cognitive system. To understand how something comes to play the role of representation, one needs to take into account those *two* facts. To say this in a different way, what is constitutive of something being a representation is both that it plays a role in guiding action *and* that, in virtue of being so tied to action, it makes representational error-detection possible.

I do not think that this argument is successful. To say that action failure indicates inaccurate or false representation, one needs to *assume* that the action in question was guided by a representation in the first place. One cannot treat action failure as way of telling that a representation is inaccurate unless one has some independent way of establishing that the action in question was guided by a representation. More broadly, it seems wrong to treat the possibility of detecting a representational error as co-constitutive of having the status of representation, because qualifying some process as "representational error detection" requires *presupposing* some understanding of representation whose being-in-error might be detected. As long as a given action is not guided by representations in the first place, it is unjustified to say that a *representational* error has been detected in this situation. At most, what has been detected is a purely *practical* error, and *not* a practical error that *resulted* from inaccurate representation, thus *indicating* the inaccuracy of a representation.

There is one last point to be made before we move on. One reviewer of this paper suggested that there is a way in which we may treat the ability for error detection as conferring a representational status on a cognitive structure. According to this suggestion, the ability to detect error is representational as long as the error detection does not modify action on-line, that is, as long as the action in question is not modified immediately. As I understand this idea, error detection should be considered representational if it is used to control action off-line; for example, if it is used to modify *future* as opposed to current actions.

I remain unconvinced. Consider two imaginary, simple artificial systems: robot A and robot B. In both cases, their movements are controlled by an artificial neural network that does not make use of representations in any recognizable or explanatorily valuable way, in the vein of Beer's aforementioned artificial agent. Now, imagine we alter both A and B by giving them the ability to receive feedback signals from the environment, which in turn enables them to detect when their respective actions are failing. There is one difference, though. In the case of robot A, the feedback signal is used to modify actions as they unfold on-line. In the case of robot B, the feedback signal is not used to modify on-line action. Rather, B uses the signal to change the connection weights in the network that controls its movement in way that will modify the way it acts in the *future*; in other words, it uses error detection off-line, to prepare future, as opposed to current, actions. If we follow the suggestion discussed here, somehow the fact that the actions being modified are "postponed" in time should make B a representational system, as opposed to a non-representational system such as A. But it seems to me that the way robot B works is not sufficiently different from the way A works to make B representational; after all, the actions of both robots were not controlled by representations in the first place. I do not see how the difference in the length of the temporal interval that separates error detection from action modification could turn a non-representational system into a representational system.

## 5. Overcoming the limitations of ACToRs. An outline of a two-factor theory of representation

Although the discussion so far has been critical of ACToRs, my aim is not to discard the action-centric approach altogether. I think ACToRs are on a right track. Rather, what I intend to argue for is that playing a role in guiding or preselecting actions is not *sufficient* to make something a representation.[5] I think that ACToRs give us a good idea of *what* it is that representations do, but they leave unanswered the question of *how* representations (as opposed to non-representations) work. Another way to put it is that action guidance is a good candidate for a *genus* of an explanatorily valuable concept of representation. What is missing from this picture is a good idea—or any idea—about the nature of *differentia specifica* that distinguishes representational action guidance from the kind of action guidance that is achieved without employing representations.

Let me illustrate this point with an example. Imagine that three people face the challenge of navigating their way from point A to point B in a city they do not know. Suppose that each person succeeds at navigating the route, but each one does it using a different strategy. Person 1 is led through the streets of the city by a local who already knows the way. Person 2 is given instructions to follow a trail composed of little red balls that, as it happens, are arranged in such a way so as to lead from A to B (thus, person 2 navigates the city by being directly coupled with what we might call, following Rick Grush, a "presentation", see Grush, 1997). Person 3 succeeds because she is given a map of the city on which both points A (her original location) and B are marked.

In each of those cases there is something—another person, a ball trail, a map—that can be attributed the role of "action-guider". But only in the case of person 3 can we justifiedly say that her action is guided by a *representation* (or at the very least *directly* so guided, given that it might be the case that person 1 is guided by a local who in turn is guided by an internal mental representation of the city). If we want a theory of representation, the crucial thing to ask is this: What distinguishes the navigational strategy employed by person 3 from the strategies employed by the two other people?

[5] It needs to be noted that in an article clarifying and defending the action guidance theory, Anderson and Anthony Chemero (2009) seem to be aware that treating action guidance as a sufficient condition for the functional role of a representation is wrong: "[…] although the guidance theory says that representations are such in virtue of their role in guiding action (all representations are action-guiders), it does not claim that everything that guides action is therefore a representation" (p. 308). However, this sort of concession raises some problems for the action guidance theory. First, it seems to suggest that the theory is only committed to what I have called in section 2 a *weak* interpretation of action-centrism. But this would be disappointing, since not many people would argue *against* the weak version of ACToRs, which simply says that representations are for action guidance. Second, this concessive move naturally raises a further question: If action guidance is by itself insufficient for making something a representation, what other conditions need to be met? This is the exact question I am attempting to answer in the present section.

I propose that the difference-making factor here is the fact that in order to succeed, person 3 exploits what we might generally call a "correspondence" between the map and the terrain. More precisely, what I think makes the case of person 3 a case of representation-use is the fact that (1) this person (a *representation user*) uses a map (a *representational vehicle*) to guide her action with respect to the terrain (*what is represented*), and (2) she does it by employing a strategy whose navigational (action-guiding) success is non-accidentally dependent on whether a certain type of *relation* holds—or holds to a sufficient *degree*—between the map (representational vehicle) and terrain (what is represented). But what sort of relation are we talking about? Without delving into detail, it may be characterized as a homomorphism or some kind of structural similarity between how constituents of the map are spatially-relationally organized and how the constituents of the terrain are spatially-relationally organized (see Bartels, 2006; Cummins, 1989; O'Brien & Opie, 2004; Swoyer, 1991).[6] Importantly, however, it needs to be stressed that the existence of this similarity is by no means *sufficient* for making the map a representation. Rather, what constitutes its representational status is *both* its action-guiding function and the way performing this function is dependent on the relevant relation holding between the map and the represented terrain.[7]

What do I mean exactly by action guidance being "dependent" on the relation between the representational vehicle and what is represented? I mean that the relation between the

[6] To be fair, it needs to be said that A&R at least (2008) do *not* deny that representational action guidance can sometimes be achieved by exploiting structural similarities, or isomorphisms, to be more precise. However, at the same time they do *not* argue that it is co-constitutive of representations that they guide actions by exploiting isomorphisms. Rather, for A&R, this is just one empirically plausible way in which representations might work or become established. This sort of claim is obviously much weaker than the two-factor theory I am arguing for here, where being a representation is *co-constituted* by the role played in action guidance by the relation between the representational vehicle and what is represented. On my view, if something is not based on exploiting such a relation, it becomes deeply problematic how it gets to play a representational role at all.

[7] It needs to be said that the choice of structural similarity as the relation that establishes the correspondence is theoretically important. As the frog example discussed in previous section shows, if we instead chose a simple co-variation or causality as the relation that holds between the vehicle and what is represented, the two-factor theory would not meet the JDC. In other words, not all correspondence-grounding relations are created equal and not all of them can serve as a basis for a workable two-factor theory of representation (see also Ramsey, 2007; von Eckhardt, 1993). By choosing structural similarity as the relation that establishes the correspondence, I treat internal representations that feature in cognitive-scientific explanations as a species of simulation-, structural- or S-representations (see Bartels, 2006; Cummins, 1989; O'Brien & Opie, 2004; Ramsey, 2007).

representation and what it represents is *causally* and thus *explanatorily* relevant for whether the representation succeeds in providing action guidance. I understand "causal relevance" along James Woodward's (2003) interventionist lines here, basically meaning to say that by intervening in whether the appropriate relation holds—or by intervening in the degree to which it holds—we could manipulate the representation user's practical or navigational success. This way, we can explain the representation-user's practical success by pointing to facts regarding whether the appropriate relation holds—or to what degree it holds—between the representational vehicle and what is represented. Take the map example again. A map is a useful action-guider only if its structure matches or resembles—or to the degree it matches or resembles—the structure of that which it represents (even when it is a highly idealized and simplified map, like the map of the London Underground). The existence (or the degree) of this resemblance is casually relevant to whether the map successfully guides action.

To sum up the discussion so far, the lesson is that representations are things that succeed in playing their action-guiding function by exploiting a certain relation between the representation itself (the vehicle) and what it represents. By no means do I think that this says anything revolutionary or new about the nature of representations. However, it seems to me that it nonetheless shows how both purely correspondence-centric (or input-centric) *and* purely action-centric (or output-centric) accounts of representation miss their target. By defending the former type of theory, we try to reduce representation to the relation between the representational vehicle and what it represents. By defending the latter type of theory, we try to reduce representation to the relationship between the representational vehicle and the representation user. But neither of those two ways of construing representations works. Problems with correspondence-centrism have already been discussed in detail in the literature (see e.g. Bickhard, 1993; Bickhard & Terveen, 1995; Miłkowski, 2013). What I have been trying to show here is that action-centrism does not succeed either, since by focusing so much on action, it fails to give us an idea about what distinguishes representational action-guidance from nonrepresentational action-guidance. However, I also think that both views describe part of the truth. I propose that to account for the nature of representations, we need a *two-factor* theory that skilfully combines both perspectives by treating representations as action-guiding structures that work by exploiting a relation between a representational vehicle and what is represented (for a somewhat similar proposal put forward in the context of teleosemantics, see

Shea, 2007, 2013).[8] I suggest that only then can we form a notion of representation that does meet the JDC and thus can explain cognitive phenomena in a genuinely and nontrivially representational way.

Note also that this two-factor theory can meet Bickhard's demand that a successful theory of representation needs to account for the representation user being able to detect a representational error. I think that the proponents of ACToRs are right in saying that in order to account for representational error detection, we need to closely tie representations to actions, so that a representation's being in error can be detected indirectly, by detecting the failure of an action that was guided by this representation. Although this idea is basically correct, by itself it is not enough, as I have shown above, to distinguish between detecting a representational error from detecting a purely practical error. To account for an ability to detect genuinely *representational* error, we need to show how the action in question—that is, an action failure upon which this detection is based—has been guided by representation in the first place. And a two-factor account tells us how we can do this: a given action was

[8] Importantly, the two-factor theory as I understand it here should be distinguished from a two-factor theory of mental content defended by philosophers like Ned Block (1986) and Hartry Field (1997). This latter theory stems from an attempt to combine two separate theories of content: the causal or covariance theory (which states that mental representation's content is determined by covariance or a causal relation between the representation and what it represents) and functional or conceptual role semantics (which states that representation's content is determined by the functional roles it plays with regards to perceptual input, other representational states, and action output). According to the theory advocated by Block and Field, intentional content is determined *both* by world-head causal or covariational relations and the functional roles played by representations in a cognitive system. There are at least three major differences between this sort of two-factor theory and the proposal I advocate in the present article. First, they differ with respect to the problem they attempt to solve. The two-factor theory I defend here is a theory of what it means for something to *function* as a representation in a cognitive system and thus explain cognitive phenomena as a representation. The two-factor theory that Block and Field opt for is a theory of mental content, that is, of how what the representation is *about* is determined. Second, although both theories are similar in that they understand one "factor" as a head-world relation, and the other as a functional property of the representational vehicle, they nonetheless differ in how they specifically construe those factors. When it comes to head-world factor, while Block and Field's theory refers to causation or covariance, the theory on offer here refers to structural similarity. When it comes to the functional factor, while Block and Field's theory refers to overall functional roles played by the representational vehicle, my two-factor theory refers solely to how the representational vehicle guides action. Third, contrary to the two-factor theory of mental content, the proposal on offer here is explicitly committed to the claim that both factors should be appropriately *related* to each other, that is, the representation's success at guiding action should depend on whether head-world relation—the structural similarity—holds (see main text for details).

representation-guided if it was guided by a structure (representational vehicle) whose success in providing action guidance non-accidentally depends on a relation between this structure itself and that with respect to which the action is guided (what is represented). Take the example I have given above. People 1, 2, and 3 can all potentially fail to get from point A to point B in an alien city. If they do not succeed, no doubt they can detect this fact. But in such a case, only person 3 has failed to achieve success after using a *representation*-based navigational strategy. Only in her case can the detection of a practical failure serve as a way of telling that a *representation* was in *error*.

One last, but crucially important thing that should be mentioned here is that maps or other representational devices that are cultural artifacts cannot, of course, be unqualifiedly treated as models for representations in a sense that could be explanatorily useful for *cognitive science*. Cognitive scientists need a notion of representation that does not require representations to be interpreted by cognitive agents with human-level cognitive capacities, according to culturally constructed and culturally transmitted rules of interpretation (O'Brien & Opie, 2004; Ramsey, 2007). Is such a notion even achievable? Answering this question in detail is way beyond the scope of this article, but I think that the short answer is yes. Take for example Rick Grush's (1997, 2004) famous emulation theory of representation. What Grush essentially proposes is an explanation of motor control according to which, in order to control the body in a swift and effective manner, the motor system employs an internal *model* of the musculoskeletal system. In line with what Grush himself claims, I think emulation theory is a genuinely representational explanation that uses the notion of internal representations as mechanical or automated "models" or "maps".

I propose that emulators owe their representational status to the fact that they meet the criteria put forward by a two-factor theory. For one thing, they meet the "action-centric" criterion, simply because they are in the business of guiding actions. As A&R put it, "[…] what, exactly, is it for R to be used as an emulator of E, that is, what ties it to E in particular? […] Our account answers this crucial question by saying that R is used as an emulator of E, and is, therefore, a representation of E, just in case (as with all representations) R is used to guide the subject's actions with respect to E […]" (2008, p. 58). This is right. But it is also *incomplete*, since it fails to take into account that a *nonrepresentational* structure could potentially guide actions with respect to E. What makes something an emulator—and thus a representation—is *how* it works, namely in a *model*- or *map*-like manner. In other words, what is co-constitutive of an emulator as a form of representation is that it meets the "correspondence-centric" criterion as well. An emulator is successful in performing its function only insofar as it generates adequate anticipations for the motor system; but it is anticipatory only insofar as the workings of an emulator "mimic" or resemble the workings of the body, or the body–world loop. For example, if we assume that emulators are "articulated"—i.e. that they are composed of elements whose dynamics of interactions are meant to model the dynamics of interactions between parts of the musculoskeletal system (see Grush 2004)—then they are successful in performing their action-guiding function to the extent to which the way the emulator works structurally or dynamically resembles the way the musculoskeletal system works. This dependence of the emulator's action-guiding function on the similarity relation is important for understanding both how emulators work and what it is exactly that gives them their representational status. Last, note that according to Grush (1997, 2004), the results of the emulation process are continuously compared against real-world sensory feedback. I propose that if the system detects a discrepancy between the two, this may count as genuine case in which the discrepancy between the world and its internal model—that is, *representational error*—has been detected (see also Miłkowski, 2013).

## 6. Conclusion

I have tried to show that while not without merits, ACToRs fall short of providing us with an explanatorily valuable notion of representation. This is because representations, as they are understood in action-centric theories, do not meet William Ramsey's job description challenge. To show this, I pointed to cognitive structures that fulfil the ACToR's functional criteria for counting as a representation, but which (1) quite manifestly do *not* play a *representational* function within a cognitive system, and thus (2) do not, and cannot explain phenomena *qua representations*. I have also put forward a diagnosis of where the problem with ACToRs lies exactly. According to this diagnosis, by putting so much emphasis on the role that representations play in controlling actions, proponents of ACToRs have lost sight of what is equally important for making representations what they are, namely the fact that using representations consists in exploiting a relation that holds between the representational vehicle and what is represented. Last, I have proposed a recipe for this theoretical situation by outlining a two-factor theory which marries the idea that representations are for guiding action with the idea that representational function depends on the existence of a correspondence (structural similarity) between representational vehicle with what is represented. Without doubt, what I have shown is just a sketch and there are a lot of questions that this two-factor theory still faces. For now however, suffice it to conclude that if I am right, then representations need two legs to stand on in order to play their explanatory role *qua* representations in cognitive science.

## References

Anderson, M. L. & Rosenberg, G. (2004). A brief introduction to action guidance theory of representation. *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, 1185–1190.

Anderson, M. L. & Rosenberg, G. (2008). Content and action: The guidance theory of representation. *Journal of Mind and Behavior, 29*, 55–86.

Anderson, M. L. & Chemero, A. (2009). Affordances and intentionality: a reply to Roberts. *Journal of Mind and Behavior, 30*, 301–312.

Bartels, A. (2006). Defending the structural concept of representation. *Theoria, 55*, 7–19.

Bechtel, W. (2008). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. London: Routledge.

Beer, R. D. (2003). The dynamics of active categorical perception in an evolved agent. *Adaptive Behavior, 11*, 209–243.

Bickhard, M. H. (1993). Representational content in humans and machines. *Journal of Experimental and Theoretical Artificial Intelligence, 5*, 285–333.

Bickhard, M. H. (1999). Interaction and representation. *Theory and Psychology, 9*, 435–458.

Bickhard, M. H. (2004a). The dynamic emergence of representation. In: H. Clapin, P. Staines & P. Slezak. (Eds.). *Representation in mind: new approaches to mental representation* (pp. 71–90). Oxford: Elsevier Science.

Bickhard, M. H. (2004b). Process and emergence: normative function and representation. *Axiomathes, 14*, 135–169.

Bickhard, M. H. (2009). The interactivist model. *Synthese, 166*, 547–591.

Bickhard, M. H., Campbell, R. L. (1996). Topologies of learning and development. *New Ideas in Psychology, 14*, 111–156.

Bickhard, M. H. & Terveen, L. (1995). *Foundational issues in artificial intelligence and cognitive science: impasse and solution*. Amsterdam: Elsevier Scientific.

Block, N. (1986). Advertisement for a semantics for psychology. *Midwest Studies in Philosophy, 10*, 615–678.

Cummins, R. (1989). *Meaning and mental representation*. Cambridge (MA): The MIT Press.

Clark, A. & Grush, R. (1999). Towards a cognitive robotics. *Adaptive Behavior, 7*, 5–16.

Dretske, F. 1988. *Explaining behavior: Reasons in a world of causes*. Cambridge (MA): The MIT Press.

Field, H. (1977). Logic, meaning, and conceptual role. *Journal of Philosophy*, 74, 379–409.

Goldman, A. (2012). A moderate approach to embodied cognitive science. *Review of Philosophy and Psychology, 3*, 71–87.

Grush, R. (1997). The architecture of representation. *Philosophical Psychology, 10*, 5–23.

Grush, R. (2004). The emulation theory of representation: motor control, imagery and perception. *Behavioral and Brain Sciences, 27*, 377–442.

Lettvin, J., Maturana, H., McCulloch, W. & Pitts, W. (1959). What the frog's eye tells the frog's brain. *Proceedings of the Institute of Radio Engineers, 47*, 1940–1951.

Miłkowski, M. (2013). *Explaining the computational mind*. Cambridge (MA): The MIT Press.

O'Brien, G. & Opie, J. (2004). Notes toward a structuralist theory of mental representation. In: H. Clapin, P. Staines & P. Slezak. (Eds.). *Representation in mind: New approaches to mental representation* (pp. 1–20). Oxford: Elsevier Science.

Ramsey, W. (2007). *Representation reconsidered*. Cambridge: Cambridge University Press.

Shapiro, L. (2010). *Embodied cognition*. New York: Routledge.

Shea, N. (2007). Consumers need information: supplementing teleosemantics with an input condition. *Philosophy and Phenomenological Research, 75*, 404–435.

Shea, N. (2013). Millikan's isomorphism requirement. In D. Ryder, J. Kingsbury & K. Williford (Eds.). *Millikan and her critics* (pp. 63–86). Oxford: Wiley-Blackwell.

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

von Eckhardt, B. (1993). *What is Cognitive Science*? Cambridge (MA): MIT Press.

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