Krystyna Bielecka

University of Białystok

Marcin Miłkowski

Institute of Philosophy and Sociology, Polish Academy of Sciences

Representationalism and Rationality: Why Mental Representation is Real

Abstract

This paper presents an argument for the realism about mechanisms, contents, and vehicles of mental representation at both the personal and subpersonal levels, and showcases its role in instrumental rationality and proper cognitive functioning. By demonstrating how misrepresentation is necessary for learning from mistakes and explaining certain failures of action, we argue that fallible rational agents must have mental representations with causally relevant vehicles of content. Our argument contributes to ongoing discussions in philosophy of mind and cognitive science by challenging anti-realist views about the nature of mental representation, and by highlighting the importance of understanding how different agents can misrepresent in pursuit of their goals. While there are potential rebuttals to our claim, our opponents must explain how agents can be rational without having mental representations. This is because mental representation is grounded in rationality.

1. Introduction

Representation and rationality are fundamentally interconnected, a connection that should be viewed through a naturalistically plausible lens. Embracing this perspective not only deepens our understanding of these concepts but also contributes to the broader project of naturalizing intentionality. Natural rationality, in its essence, involves the occasional misrepresentation: no natural rational being is omniscient, and even the most judicious among us make inadvertent mistakes. These misrepresentations, paradoxically, underline our capacity to represent. Thus, viewing us as rational beings necessitates viewing us in representational terms.

This claim warrants some clarification. We are not suggesting that misrepresentation is chronologically, developmentally, or evolutionarily prior to representation. Rather, our simple assertion is that the necessity to posit misrepresentation logically presupposes the existence of prior representation capacities.

Let us situate our view in the debate over the role of misrepresentation in naturalizing intentionality. Theories of mental representation must not only account for the possibility of misrepresentation (Dretske 1986). They can also be strengthened by relying on a causal understanding of error-detection, as it bolsters realism about contents. The ability to detect errors serves as evidence not only for the functionality of one's cognitive abilities, but also for understanding (and determining) the contents possessed by representational vehicles.

Jerry Fodor emphasized that the ability to detect an error in a representation was evidence that this content was part of one’s psychology. This is evident in his discussion of

the difference between human conceptual abilities and a frog's ability to detect when it has mistakenly targeted an object:

> Sometimes Macbeth starts at mere dagger appearances; but most of the time he startles only if there's a dagger. What Macbeth and I have in common—and what distinguishes our case from the frog's—is that *though he and I both make mistakes, we are both in a position to recover*. By contrast, frogs *have no way at all* of telling lies from bee-bees (Fodor 1992, p. 107).

Fodor, however, did not extensively explore this idea. In contrast, Mark Bickhard has made error detection a criterion for the adequacy of any naturalistic approach to intentionality (Bickhard 1993, 2009). According to his account, error detection occurs in basic representations that drive the actions of an agent, provided the agent can recognize when anticipated actions fail. Therefore, the ability to detect failure is something that frogs may lack, at least in regard to things they snap at (if Fodor's assessment is correct), but it is a capability that we (and Macbeth) possess.

Earlier attempts to underscore the importance of error-correction mechanisms in representational theories have provided a clear perspective on the causal role of content in guiding actions (Bickhard 1993, 2009; Ryder 2004; Lee 2019; Bielecka and Miłkowski 2020; Buckner 2022). In this paper, we argue that finite rational agents are prone to errors due to misrepresentation. Unlike previous naturalistic approaches to intentionality, which often relied on rationality assumptions but avoided realism (Davidson 1984; Dennett 1987; Molder 2010), we strive to ground mental representation in instrumental rationality, thereby establishing a realistic perspective. In essence, our argument grounds representation in rationality.

The structure of this paper is as follows: We begin by stating our general assumptions and then sequentially develop our arguments in two distinct phases. Our first argument implies representation at the personal level. However, in the succeeding phase, we maintain that representation can be extended to any spatiotemporal scale, encompassing even subpersonal entities. The initial argument demonstrates the necessity of misrepresentation in elucidating why an agent can remain instrumentally rational despite performing an unsuccessful action. The subsequent argument elaborates on how such misrepresentation should be understood as having a subpersonal vehicle and mechanisms.

In the penultimate section, we briefly review potential rebuttals. Next, we argue that those who disagree with representationalism bear the burden of explaining how agents can be rational without having mental representations. Finally, we conclude by summarizing our main argument. Our argument shows that mental representations are necessary for instrumental rationality and must have causally relevant mechanisms, contents, and vehicles. While there are potential rebuttals to this claim, our opponents must explain how agents can be rational without having mental representations. Therefore, we aim to dispense with the anti-realism about mechanisms, contents, and vehicles of mental representation.

2. Explaining Failures of Action with Misrepresentation

To clarify our terminology, by 'agent' we mean any entity that is capable of pursuing goals or taking actions to achieve desired outcomes. Rational agents, a specific type of agent, choose means that are conducive to their established goals. In other words, we assume that some agents are instrumentally rational, which means that they can select appropriate means to achieve their goals. This assumption is backed by ample evidence of humans consistently

selecting means that align with their goals, even while making errors that highlight their lack of omniscience. We do not, however, stipulate that instrumentally rational actions must stem from semantically rational processes or deliberative planning. Furthermore, we do not insist that agents must maximize expected utility through apt choices to qualify as instrumentally rational. A more plausible assumption is that most agents are boundedly rational, with their decision-making processes significantly shaped and constrained by limits on their capacities and complexity (Simon 1956, p. 129). Lastly, we do not presuppose that instrumental rationality necessitates explicit reasoning or conscious awareness, thereby allowing for habitual or instinctual actions to be classified as instrumentally rational, regardless of whether they result from deliberation.

Instead of presupposing that rationality relies on representation, our aim is to demonstrate that instrumental rationality requires particular kinds of causal-computational mechanisms. Our argument is constructed in the vein of computationalism (for recent defenses, see Piccinini 2020; Colombo and Piccinini 2023; Fresco 2014; Miłkowski 2013), as it's the preferred viewpoint of the majority of representationalism advocates. As Coelho Mollo (2021) argues, when considered in non-semantic mechanistic terms, computationalism, backed by teleomechanistic considerations, helps to explain and naturalize representation. Moreover, we deem the objections raised against the computational perspective to be unconvincing (for a review, see Miłkowski 2018).[1]

[1] Simultaneously, there exist non-computational perspectives on the mind, such as those inspired by ecological psychology, that greatly depend on the concept of mental representation. For instance, consider Mark Bickhard's (Bickhard 1993, 2009) interactivist model. However, the applicability of our argument to non-representational dynamical views of the mind falls outside the purview of our paper, despite our optimism regarding this possibility.

Our objective is to establish that mental representations are indispensable constituents of causal-computational mechanisms in natural rational agents. This positions our argument against both ontological and explanatory anti-realism about mental representation. We adopt attributes associated with mental representation by anti-realists without endorsing any specific conception of mental representation beyond a fairly bland assumption that they have teleofunctional characteristics. Thus, we don't commit to any particular formats of vehicles of mental representation, an issue recently debated by Mollo and Vernazzani (2023) and Yousif (2022), leaving this issue open. Instead, our focus is on the semantic features of representation, which requires that vehicles exist and are operational but does not decide their specific formats. We concur with Ramsey (2023), however, that vehicles are indispensable in a conception of mental representation for its satisfactory defense.

Following radical enactivists, we conceive of mental representations as possessing contents, defined in terms of satisfaction conditions (Hutto and Myin 2013). In short, we assume the following definitions:

- (R1) *A representation vehicle* is the physical medium of information that is processed by causal-computational representational mechanisms.
- (R2) *A representational mechanism* is the causal-computational mechanism that operates on representational vehicles and engages in representing processes when functioning correctly.
- (R3) A representation *target* is the (possibly vacuous) referent of representing processes.

- (R4) Representation *contents* are the satisfaction conditions of representational vehicles, a characteristic that, in our perspective, does not supervene locally on the structure of the agent alone, therefore making our position externalist.[2]

Additionally, we believe the most controversial property for anti-representationalist is that representation vehicles and mechanisms can be subpersonal; most of them would not deny the existence of public representations and their physical vehicles or contents. This is why we aim to establish that mental representations depend on subpersonal processing, which include subpersonal representational mechanisms, vehicles, and contents. Thus, we first establish that we must attribute mental representation in terms of contents to rational agents, and then demonstrate that these contents have associated subpersonal vehicles and mechanisms, possibly also with specifically subpersonal contents.

Agents may fail to achieve their goals for various reasons, such as choosing means inappropriate to their ends, unexpected external factors disrupting the causal chain between their actions and goals (e.g., a meteor hitting the Earth), or misrepresenting the available means in a given situation. When an agent consistently uses ineffective means to achieve their goals, we may view them as irrational in their actions. However, unexpected events or misrepresentation are not always avoidable, and the agent may remain entirely rational while pursuing an action that ultimately proves ineffective. The key point is that misrepresentation, a common cause of goal-failure, is sometimes the only thing that prevents us from concluding that an agent is irrational.

Here is the summary of the argument:

[2] We do not assume semantic externalism in our arguments below, however.

1. Agent A is instrumentally rational.
2. If A is instrumentally rational and represents A's situation in a sufficiently adequate manner, A selects means that are likely to lead to A's goals.
3. But there are cases when A selects means that are not likely to lead to A's goals.
4. Thus, A is either not instrumentally rational or does not represent A's situation in a sufficiently adequate manner (2, 3, modus tollens)
5. Thus, A does not represent A's situation in a sufficiently appropriate manner (1, 4, disjunctive syllogism).

This argument is logically valid.[3] We also believe it to be sound for many agents. For example, consider the following scenario: Alice, a university student, plans to catch the direct bus from a stop that is simply a pole, with no posted schedule. She relies on her memory that the bus usually arrives there at 9:05 AM. However, unbeknownst to her, the bus service was canceled the previous day. As a result, waiting for the bus will not get her to the university. Without the ability to ascribe to Alice the false belief that there is a direct bus from that bus stop to the university (which would be rejected by those who reject the notion of mental representation), we might conclude that Alice is being irrational by sitting on the bench waiting for a bus that will not come. However, for all we know, Alice is instrumentally rational and simply unaware of the cancellation. In this case, we would need to explain Alice's failure to reach the university in some other way, rather than attributing it to irrationality.

Therefore, if Bob wants to explain why Alice's action failed to achieve her goal, he should attribute it to misrepresentation. In this scenario, the misrepresentation is attributed

[3] We interpret Premise 2 as an instance of material implication. Consequently, the antecedent of the conditional is regarded as a sufficient condition for the truth of the consequent.

to Alice. Consequently, we have good reason to consider misrepresentation as the root of such failures.

Note that the argument does not require the misrepresentation to be tied to linguistic abilities. To further illustrate our point about misrepresentation in instrumental rationality, consider another example involving a dog. In this case, the dog may be searching for a bone that was previously stored in the garden, but is unaware that the owner has already retrieved it. This example highlights how misrepresentation can occur even in non-human animals and is not limited to human cognition.[4]

At this point, our opponent may argue that our argument demonstrates only that one must ascribe contents to instrumentally rational agents, but that one can still be instrumental about the vehicles of such contents, in particular when subpersonal mechanisms are concerned. In the next section, we will argue against this move.

3. Mental Representation Going Subpersonal

As we proceed to the subpersonal in the second part of our argument, we adhere to Daniel Dennett's original conception of subpersonal theories. As Dennett understands subpersonal theories, they "proceed by analyzing a person into an organization of subsystems" (Dennett 1978, p. 154; see also Drayson 2014). This view establishes a part-whole relationship between

[4] Donald Davidson (1982) famously denies representational capacities to non-linguistic animals, a view critiqued as potentially question-begging and over-intellectualized by Tyler Burge (2010). In contrast, comparative psychology provides ample evidence of animals' capability for practical inference. Bence Nanay refers to this as 'pragmatic inference' and aligns with Burge's perspective. Nanay's theory would ascribe 'pragmatic representations' to the dog seeking for the bone (Nanay 2013). However, terminological and semantic disputes persist, complicating the issue of whether these capacities in animals are truly underwritten by representation. For an exploration of possible criteria of ascribing mental representations to animals, in comparative psychology, see Buckner's (2014).

the personal and the subpersonal, a perspective we maintain in contrast to others who have diverged from Dennett's interpretation.

In this context, our use of 'subpersonal' is not meant to imply the autonomy of the personal level nor does it necessitate the lack of introspective access as a defining condition (see Rupert 2023 for a comprehensive review). Nonetheless, we posit that the limited introspective access to processes like speech generation suggests the presence of subpersonal mechanisms. Consequently, it lends credibility to hypotheses about subpersonal mechanisms. The actual existence, organization, and operation of these mechanisms warrant further studies, in line with the recommendations of proponents of the new mechanistic account of explanation (Craver 2007; Machamer et al. 2000; Piccinini 2020).

Moreover, the same features may be posited at both the personal and subpersonal levels. This is exactly what is at stake here. There is a way to generalize the previous argument by noting that instrumentally rational agents adapt in order to achieve their goals, often by learning from their mistakes. To explain how they adapt and learn, cognitive (neuro)science commonly appeals to subpersonal mechanisms that operate upon their representations, which we understand as involving processing vehicles of semantic information and responding to their semantic, rather than merely syntactic, properties.

We argue that at least some success or failures can be attributed to subpersonal learning processes, rather than personal-level ones, especially when the learning process happens without the agent's awareness. This suggests the involvement of subpersonal representational mechanisms. There also exist subpersonal processes that explain the agent's action, and these processes can function under the agent's conscious control. However, the point of our argument is to demonstrate that these two types of processes could be dissociated.

The crucial consideration is that there is a strong conceptual connection between the success of an action and the agent's representational accuracy. While not all success or failure of biological agents requires representation, in particular for simple tracking behaviors such as chemotaxis or phonotaxis (Burge 2010), learning from mistakes requires at least rudimentary forms of representation. This is because learning explains the increasing success in achieving appropriate goals (or serving appropriate functions) by demonstrating an increasing *accuracy* of the representation.

The accuracy of representation is "a fuel for success" of one's action (Godfrey-Smith 1996, pp. 171–195). Supporters of "success-linked semantics" argue that representational accuracy has a causal influence on the success of an action (Shea 2018). For example, the degree of similarity between structural representations and their targets is relevant to the degree of success, even if there is a complex trade-off between "representation's structural complexity and the temporal or computational resources (costs) that real-life cognitive systems have at their disposal" (Gładziejewski and Miłkowski 2017, p. 343). This implies that the failure of one's action can be explained by representational inaccuracy, in particular when no other explanations screen off the representational one.[5]

The connection we are exploring is conceptual, and we believe it's plausible even if there is little evidence that reliance on functional considerations of success semantics influences explanatory practices in neuroscience (Favela and Machery 2023).[6] Our inspiration

[5] We do not mean to suggest that representational accuracy is the only relevant factor in the success or failure of one's actions. In fact, this couldn't be further from the truth. For example, we can have a perfect understanding of the causes of climate change, but have very little, if any, practical means to halt it. Our argument holds only for scenarios in which the agent (or a device) can act in accordance with the contents of the representation. Of course, their actions can also be (sometimes) successful without representation.

[6] While we champion naturalism, we acknowledge the clear fallacy in deriving normative advice directly from the description of scientific practice, particularly when focusing on a single, albeit vast

is rooted in, and is an integral part of, control theory. The Good Regulator Theorem states that all good (optimal) controllers have models of whatever they control (Conant and Ashby 1970). To control all aspects of entity E, the controller must have a model M with as many degrees of freedom as are inherent in E. If there is no information about E in model M available to controller C, we cannot explain C's success. Inaccuracy in M can also explain its failure.

In this context, instrumental rationality can be seen as an example of good (enough) control, and wherever we find good control systems flexibly adapting to various conditions, we should expect representation. Specifically, in rapidly changing circumstances where entities being controlled may undergo significant changes, controllers should be able to adapt their models by detecting that their accuracy was insufficient. Similarly, poor control should lead to revising one's models. Nevertheless, some control systems may adapt by revising their models through a negative feedback mechanism, which need *not* involve sensitivity to the model's contents. In more complex forms of control, the model's accuracy can be monitored by checking its consistency with multiple sources of feedback, which implies sensitivity to the accuracy of the model (Bielecka and Miłkowski 2020). To sum up, the conceptual connection becomes even stronger for adaptive model-based control, which involves learning from mistakes in complex monitoring scenarios.

We propose that subpersonal cognitive processes in humans are involved in adaptive model-based control, which seems uncontroversial as far as there are multiple lines of empirical evidence regarding our metacognitive processing. Our goal here is to demonstrate that there could be subpersonal representational processes involved in adaptive model-based

---

and diverse, field of inquiry such as neuroscience. This is especially pertinent when this field may itself require conceptual engineering. Despite the apparent success of the representational research program (Bechtel 2016; Thomson and Piccinini 2018), skepticism regarding the concept of mental representations remains widespread among neuroscientists (Brette 2019).

control that occur without our personal-level introspective access or beliefs. This is sufficient to show that misrepresentation can occur at the subpersonal level without necessarily being present at the personal level. Notice the limitation of our argument: we do not advocate for any particular format of vehicles, as our argument is orthogonal to this issue. Instead, we focus on the representational function of subpersonal and personal mechanisms, which requires physical vehicles, but does not (fully) determine their format.

Some subpersonal learning processes occur without our awareness. For example, Bob could be a psychological subject who adapts his actions without any introspective access, such as responding to masked signals by inhibiting certain action patterns (Lau and Passingham 2007; van Gaal et al. 2011). Cognitive control does not require the agent to be aware that it is happening. In such a case, the agent's behavior is changed, but the agent seems completely unaware of this. The behavior could still be instrumentally rational and biologically functional, since over time it could minimize the discrepancy between the actual result of the action and the overall goal. Given that we have experimental evidence (in neuroscience) of the mechanisms of cognitive control that are responsible for such adaptation of unconscious action, we have reason to believe that their operation cannot proceed without correcting misrepresentations about the specific circumstances that determine the action's success. In other words, there is evidence that learning from error can occur without our conscious access, and the occurrence of such learning processes is evidence that our subpersonal processes are also instrumentally rational.[7]

Now consider a situation in which Bob initially fails to perform an action and cannot explain why. If his failures continue (for example, because his subpersonal processes are

[7] This also implies that subpersonal processes can enter "the space of reasons", in contrast to the Sellarsian claims to the contrary (Drayson 2014).

impaired in some way), we can blame Bob's cognitive control mechanisms. However, it would be inappropriate to blame Bob as a person, since he may never have been fully aware of the masked signal. Therefore, we must posit such subpersonal mechanisms to understand his action failure. At the same time, if Bob eventually succeeds, the success is due to his subpersonal machinery, which corrects the action by inhibiting a certain action pattern based on a previously misrepresented experimental situation.

In short, the argument in this section is:

1. Agent A's subpersonal mechanisms function properly.
2. If agent A performs an action without any introspective access to critical features of action success conditions, A's success or failure is attributable only to A's subpersonal mechanism.
3. Agent A performs an action without any introspective access to critical features of its success conditions.
4. If A's subpersonal mechanisms function properly and represent A's situation in a sufficiently adequate manner, they select means that are likely to lead to A's goals.
5. But there are cases when subpersonal mechanisms select means that are not likely to lead to A's goals.
6. Thus, the subpersonal mechanism does not function properly or does not represent A's situation in a sufficiently adequate manner. (4, 5, modus tollens)
7. Thus, the subpersonal mechanism does not represent A's situation in a sufficiently adequate manner. (1, 6, disjunctive syllogism)

The difference between this argument and the previous one is that we assume that instrumental rationality is supported by properly functioning subpersonal mechanisms, which contribute, among other things, to learning from our mistakes over time (premise 1). A's subpersonal mechanisms typically contribute to achieving their personal goals, which is why A's successes and failures can be (partially) explained by recourse to these mechanisms. Admittedly, biological mechanisms that are functional in one regard can sometimes inhibit an A's ability to achieve their goals in other ways. For example, a sweet tooth can make it difficult for A to maintain a healthy weight. In such cases, these mechanisms do not explain how A achieves their avowed goals (healthy weight); instead, they simply serve their biological function of seeking high concentrations of sugar in the food.[8]

It's important to clarify what we mean by the 'proper functioning' of one's subpersonal mechanisms.[9] In the realm of logic, the correctness of arguments can be assessed in two ways: formally, as validity, which is independent of the truth of the premises, or more stringently, as soundness, which requires the premises to be true. Broadening this to computational systems, miscomputation can encompass errors of validity (termed 'conceptual' by Fresco and Primiero), as well as material errors, which involve the general interaction of the computational system under normal conditions, and physical errors stemming from hardware malfunctions (Fresco and Primiero 2013). When an action fails due to misrepresentation, we're dealing with valid computation but violations of material conditions. Physical damage might also be a factor, such as in the case of brain lesions, but in

[8] While the structure of biological functions served by various mechanisms in a biological individual can be analyzed in terms of a web of interdependencies, a full examination of this complex structure and subpersonal communication processes falls outside the scope of this paper. Suffice it to say that our view implies that not only brains but also biological individuals and distributed systems are massively representational, or contain enormous numbers of representations (see also Rupert 2011).

[9] We thank an anonymous reviewer for pressing us on this issue.

such cases, we wouldn't attribute the action's failure to misrepresentation. To put it succinctly, when we say that subpersonal mechanisms function properly, we mean that they compute the functions they are supposed to compute in normal conditions, thus, displaying no errors of validity.[10]

In essence, the capacity for learning from mistakes, which implies sensitivity to satisfaction conditions of contents, necessitates mental representation, which is integral to instrumental rationality and optimal cognitive functioning. However, this doesn't imply that machine learning *products*, such as artificial neural networks, inherently possess intentionality. While the *process* of machine learning fundamentally involves computing a discrepancy between expected and actual values derived from an inferential process on a data structure, it doesn't necessarily mean that the data structure is genuinely representational. Consequently, machine learning doesn't directly address the issue of naturalizing intentionality or the symbol grounding problem (Harnad 1990).

Similarly, control based on negative feedback doesn't inherently generate contentful representations. As we argued (Bielecka and Miłkowski 2020), discrepancy detection may remain "non-semantic" (p. 302). A device like the Watt governor, while measuring and controlling, doesn't function as a representational mechanism, even though its components may carry semantic information about engine speed (p. 303). This is because not *all* discrepancy detection is semantically relevant. It becomes so only if the detected discrepancy influences a downstream mechanism, which is sensitive not just to the (proximal) physical attributes of the representational vehicle, but also to its (distal) satisfaction conditions. To

[10] In this paper, we only assume that representational mechanisms are functional (Garson 2013), remaining agnostic about the notion of function that might be suitable in this context, as there is an ongoing debate on this matter (Dewhurst 2018; Miłkowski 2013; Piccinini 2015).

fulfill this criterion, the downstream mechanism must track the accuracy of the supposed vehicle in some way. A basic negative-feedback controller, like the Watt governor, ignores potential deviations of the detected discrepancy from the appropriate or expected engine speed. However, a more advanced controller could compare two or more independent speed indicators (whereby accuracy may be tracked) and adjust the speed if necessary.

In fact, in many trained artificial neural networks (unlike machine learning processes), discrepancy detection is often entirely absent since the network remains unchanged over time when used for downstream inference. The training system might be sensitive to semantic features if discrepancy detection is implemented accordingly, typically via a cost function, but this isn't always the case. Yet, if the network is constantly updating its values, as seen in predictive coding setups, it could track accuracy information, as recently suggested by Buckner (2022). Moreover, as Buckner argues, sensitivity to error in complex cognitive architectures provides evidence for ascribing determinate contents to subpersonal vehicles (a very similar point is found in Miłkowski 2015).

To sum up, Watt's governor malfunction is hardly representational. It is merely functional in tracking the engine speed, but tracking is insufficient for representation. However, some of our subpersonal mechanisms can misrepresent, if only they respond to semantic accuracy of contents of their vehicles. This is the specific difference that distinguishes representational malfunction from malfunction *simpliciter*.[11]

We propose therefore that the notion of error and learning can be applied not only at the personal level, but also at subpersonal levels. This means that this notion, along with the notion of misrepresentation, is level- or scale-free. Therefore, it is reasonable to suspect that

[11] We thank the anonymous reviewer for pressing us on this point.

subpersonal representational mechanisms are also operating when Bob is fully aware of what caused him to change his course of action. They explain why he was able to do so.

4. Objections and Responses

1. *Human beings are not fully instrumentally rational.* Thus, the argument is unsound.

While it is true that people may sometimes be delusional, engage in self-deception, or fail to achieve their stated goals, such as quitting smoking, we are only assuming that people are sometimes able to achieve their goals through proper selection of means, and that misrepresentation rather than irrationality may sometimes be the cause of their choice of means. Hence, this objection does not address our argument. We do not assume that people are ideally rational, but we do believe that there are finite rational agents (otherwise, our argument would be unsound). Additionally, the functioning of certain subpersonal mechanisms may actually explain why some irrational behaviors are resistant to change.

2. *The definition of rationality in terms of making choices begs the question*.

Additionally, it could be argued that defining rationality in terms of making choices assumes that one's choices or goals are explicitly represented, when the decision-making process does not need to be understood in this way. To address this, we must emphasize that we are following the traditional understanding of instrumental rationality, but we do not require that choices be explicitly represented. It is sufficient that they are displayed through behavior.

3. *This argument does not prove that agents are representational; it demonstrates that we treat them as if they were. However, this practice is only socionormative.*

The first argument only establishes personal-level representations, which could be understood in a socionormative way: as the ascribed contents of linguistic representations. However, while we note that animals are also attributed with non-linguistic representations, Section 3 establishes that there are subpersonal representations in agents capable of learning from their own mistakes. Both of these cannot be merely socionormative. First, solitary animals also engage in such learning, lacking our "form of life" or "The Background" or any other form of social context one might ascribe to people. Second, our subpersonal processes can often operate without any external help, and it is unlikely that socionormative influences have significant impact on Bob's subpersonal learning processes in the sketched scenario. Moreover, there is no evidence that socionormative practices are needed for perception or solitary thought, as emphasized by Burge (2010, p. 269) in his critique of Davidson.

4. *Instrumental rationality is itself only an instrumental construct (forgive the pun).*

One could argue that there is no evidence that people are truly instrumentally rational. We could simply adopt an intentional stance towards them, treating them as if they were instrumentally rational.

While this approach might be appealing to anti-representationalists, it's worth noting that the intentional stance, along with Davidson's principle of charity (Davidson 1973) traditionally necessitates ascribing true contents (McGeer 1992). Our critics may argue that

this framework can accommodate a nuanced ascription of false contents, thus acknowledging misrepresentation and error.

This argument is not without merit, yet it lacks a comprehensive account of ascribing misrepresentation to a rational agent within this framework. Dennett (1987, p. 103) himself acknowledges cognitive mistakes leave an 'uninterpretable gap' in the narrative from the intentional stance. To defend his conception, Dennett appeals to general indeterminacy considerations, and claims that there are no facts of the matter that could decide the issue. But even if his considerations about folk psychology and belief attribution might seem plausible, increasing the indeterminacy of content ascription is a bug rather than a feature of a theory of representation. In our bus stop scenario, the most parsimonious story about Alice's *unintended* failure to get to the university is the one that mentions the falsity of what she remembered to be the time when the bus previously arrived. One can concoct other stories, but the point should not be to defend indeterminacy, but to admit that rational agents need not be omniscient, and that their rationality need not be arbitrarily idealized.

From the instrumentalist perspective, it is much easier to ascribe *intended* false representations, such as the ones involved in pretense play or lying. This is because the target of our belief ascription has a true belief that what they do is mere pretense or lie, so typical charity considerations apply. For this reason, one can also attempt to account for human pretense play in terms socionormative practices as well (Weichold and Rucińska 2022), as this kind of activity actually succeeds and does not undermine our overall reliance on rationality considerations. If a kid successfully pretends to be a T. Rex, we have little reason to suspect any problems with their instrumental rationality.

However, attributing *unintended* errors to an agent presents a challenge. Davidson (2004, p. 141) insists that to truly err, the creature must recognize the error. But as McGeer

notes, it's unclear how one can maximize rationality and coherence of belief when an agent holds a false belief.

Here, we flip the script: in our perspective, ascribing false contents is what maximizes rationality. If an agent can subsequently identify incoherence in their web of belief, we can attribute the capacity for error detection to them. This capacity, rather than maintaining a fully coherent web of belief—a task that is computationally intractable (Thagard 2000; Zawidzki 2013; Zeppi and Blokpoel 2017)—upholds instrumental rationality.

As far as unintended misrepresentation is concerned, instrumentalism has found no answer how to systematically ascribe it, in particular in a mechanistically plausible manner. Lacking computationally feasible and biologically plausible mechanistic realizations, the principle of charity, intentional stance, and rational interpretation may not be psychological processes but mere philosophical rational reconstructions that may not mirror our cognitive abilities at all. If an instrumentalist is willing to embrace the massive indeterminacy and computational intractability of their proposal, our argument may not persuade them. But we would rather be instrumentalists about philosophical rational reconstructions.

One could also claim that we only ascribe learning instrumentally. However, it seems natural that we would also adopt this approach towards other properties; representations are not unique in this regard. Our argument is only compelling for those who believe that at least some everyday instances of instrumental rationality and learning are not just theoretically useful constructs, but can be treated realistically because we can intervene in the process of selecting appropriate means or in the process of learning. Of course, this assumption could be rejected. We do not intend to challenge all forms of antirealism in this paper.

5. *Isn't rationality possible without subpersonal representation?*

In this paper, we understand instrumental rationality as not requiring learning through trial and error or learning from one's own mistakes. We did not argue that subpersonal representation is necessary for personal-level rationality. Some agents may rely on simple magnitude tracking, such as following glucose gradients in water. In our argument, subpersonal representation is only established for agents that are able to succeed without the agent's awareness, which may occur during their learning from their own mistakes. This argument highlights a dissociation between personal and subpersonal forms of representation. While we have evidence for this dissociation, we do not need to assume that it is typical in order to establish the existence of subpersonal representation as distinct from the personal one, which is all we aim to do in Section 3. In fact, we do not assume that the concept of 'subpersonal' is defined by a lack of awareness. However, we believe this lack provides significant evidence supporting the hypothesis that subpersonal processing may be involved. While this evidence is insufficient to establish computational and causal details, it is adequate to lend credibility to a representational hypothesis.

6. *Isn't this just a recycled argument from illusion?*

Not so. The argument from illusion was used to argue for the existence of sense data: from non-veridical perception to veridical acquaintance of sense-data (Ayer 1940). We are not making a claim about the existence (or non-existence) of sense data. Vehicles of misrepresentation are not only physical, unlike sense data, but they also bear non-veridical contents (whereas sense data is by definition veridical).

Our argument focuses on the concept of misrepresentation, which is distinct from the concept of illusion. In contrast to the argument from illusion, it is easy to recover from the error in our argument by correcting the misrepresentation. It is not possible to recover from the error assumed in the argument from illusion without damaging the perceptual system (one cannot normally stop seeing the stick in the water as bent). Our argument posits the existence of misrepresentation as a means of vindicating the agent's instrumental rationality and proper cognitive functioning: The error in action is explained by the error in representation.

7. Do you suggest that you could ascribe (subpersonal) representation from the armchair?

One could object by saying that we are free to posit subpersonal representation for any successful behavior, regardless of whether it is actually involved, given our claim that there is a conceptual connection between the success of an action and representational accuracy. While it is true that conceptual connections can indicate relationships that can be established without further empirical evidence, our arguments in Sections 2 and 3 rely on the truth of their premise 1. Even if we may attribute instrumental rationality to any agent, some agents fail to exhibit this rationality to a greater or lesser extent. For example, an agent may be delusional, acting on their psychotic states, or intoxicated. Empirical evidence is also required in order to establish that a device is functioning properly. In fact, even establishing the function of a device requires substantial empirical evidence (for example, it may be necessary to examine the evolutionary histories of biological mechanisms).

The same logic applies to the Good Regulator Theorem. Good regulators require models, but there may be worse regulators that still do their job in certain circumstances. We

can only conclude that these regulators rely on models (understood along the lines of control theory) if we establish that their control is optimal. As we note, this is still insufficient to establish that they are representational unless they process error information in a way that respects its semantic content (see (Bielecka and Miłkowski 2020) for a deeper study of error detection and representation).

5. Conclusion

In this paper, we argued that a class of action failures can be explained by inaccurate representation. Instead of attributing these failures to irrationality, we can vindicate rationality by positing misrepresentation. In doing so, we also posit the existence of mental representation—both at the personal and subpersonal levels—as a necessary condition for instrumental rationality and proper cognitive functioning.

We addressed objections to this view and argued that the burden is now on anti-representationalists to provide an alternative account of how agents can remain rational without mental contents. A tall order, we think. Despite the variety of defenses for anti-representationalism provided in various works (Chemero 2000; Degenaar and Myin 2014; Downey 2018; Facchin 2021; Garzón 2008; Hutto and Myin 2013; Kohár 2023; Orlandi 2014; Raja 2018; Van Gelder 1995) our argument challenges both its explanatory and ontological versions. We show that misrepresentation is important in explaining action failures, contradicting the main claim of explanatory anti-representationalism that content and vehicles of mental representations are explanatorily irrelevant. Furthermore, we argue against ontological anti-representationalism by demonstrating that the semantic features of mental representations are causally relevant to rational action. A detailed discussion of all anti-

representational positions is beyond the scope of this paper, but we firmly believe that a realistic approach to instrumental rationality in finite cognitive agents requires a realistic view of the content and vehicles of mental representation, both personal and subpersonal.

Every philosophical argument has its assumptions, and ours is no different. These assumptions, of course, limit the extent of our claims. We have assumed the existence of fallible rational beings, a premise that one could reject to claim our argument unsound. While we think it's unlikely that there are any infallible rational beings, or that there are no rational beings at all, we acknowledge that some philosophers might defend these positions, however controversial they might be. We've also assumed the computational view of the mind to make our points, which could also be a point of contention for some. Finally, we have not made any general argument against instrumentalism as a whole. We accept this limitation.

Acknowledgements

The authors wish to thank Paweł Gładziejewski, Tomasz Korbak, Wojciech Mamak, Daniel Piecka, and Wiktor Rorot, as well as five anonymous reviewers of this journal for their extensive comments to the earlier draft of this paper.

Funding:

The work on this paper was funded by National Science Center from research project 2016/23 D/HS1/02205 (PI: Krystyna Bielecka)

Conflict of interest

None

References

Ayer, A. J. (1940). *The foundations of empirical knowledge*. New York: Macmillan.

Bechtel, W. (2016). Investigating neural representations: the tale of place cells. *Synthese*, *193*(5), 1287–1321. https://doi.org/10.1007/s11229-014-0480-8

Bickhard, M. H. (1993). Representational content in humans and machines. *Journal of Experimental & Theoretical Artificial Intelligence*, *5*(4), 285–333. https://doi.org/10.1080/09528139308953775

Bickhard, M. H. (2009). The interactivist model. *Synthese*, *166*(3), 547–591. https://doi.org/10.1007/s11229-008-9375-x

Bielecka, K., & Miłkowski, M. (2020). Error Detection and Representational Mechanisms. In J. Smortchkova, K. Dołęga, & T. Schlicht (Eds.), *What are Mental Representations?* (pp. 287–313). New York: Oxford University Press.

Brette, R. (2019). Is coding a relevant metaphor for the brain? *Behavioral and Brain Sciences*, *42*, E215. https://doi.org/10.1017/S0140525X19000049

Buckner, C. (2014). The Semantic Problem(s) with Research on Animal Mind-Reading. *Mind & Language*, *29*(5), 566–589. https://doi.org/10.1111/mila.12066

Buckner, C. (2022). A Forward-Looking Theory of Content. *Ergo: an Open Access Journal of Philosophy*, *8*, 37. https://doi.org/10.3998/ergo.2238

Burge, T. (2010). *Origins of objectivity*. Oxford: Oxford University Press.

Chemero, A. (2000). Anti-Representationalism and the Dynamical Stance. *Philosophy of Science*, *67*(4), 625–647. https://doi.org/10.1086/392858

Coelho Mollo, D. (2021). Why go for a computation-based approach to cognitive representation. *Synthese*, *199*, 6875–6895. https://doi.org/10.1007/s11229-021-03097-5

Coelho Mollo, D., & Vernazzani, A. (2023). The Formats of Cognitive Representation: A Computational Account. *Philosophy of Science*, 1–20. https://doi.org/10.1017/psa.2023.123

Colombo, M., & Piccinini, G. (2023). *The Computational Theory of Mind*. Cambridge: Cambridge University Press. https://www.cambridge.org/core/elements/computational-theory-of-mind/A56A0340AD1954C258EF6962AF450900. Accessed 12 January 2024

Conant, R. C., & Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, *1*(2), 89–97. https://doi.org/10.1080/00207727008920220

Craver, C. F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Oxford University Press.

Davidson, D. (1973). On the Very Idea of a Conceptual Scheme. *Proceedings and Addresses of the American Philosophical Association*, *47*, 5–20. https://doi.org/10.2307/3129898

Davidson, D. (1982). Rational Animals. *Dialectica*, *36*(4), 317–327. https://doi.org/10.1111/j.1746-8361.1982.tb01546.x

Davidson, D. (1984). *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press.

Davidson, D. (2004). *Problems of rationality*. Oxford; New York: Clarendon Press; Oxford University Press.

Degenaar, J., & Myin, E. (2014). Representation-hunger reconsidered. *Synthese*, *191*(15), 3639–3648. https://doi.org/10.1007/s11229-014-0484-4

Dennett, D. C. (1978). *Brainstorms. Philosophical Essays on Mind and Psychology*. Cambridge, Mass.: MIT Press.

Dennett, D. C. (1987). *The Intentional Stance*. Cambridge, Mass.: MIT Press.

Dewhurst, J. (2018). Computing Mechanisms Without Proper Functions. *Minds and Machines*. https://doi.org/10/gd3274

Downey, A. (2018). Predictive processing and the representation wars: a victory for the eliminativist (via fictionalism). *Synthese*, *195*(12), 5115–5139. https://doi.org/10.1007/s11229-017-1442-8

Drayson, Z. (2014). The Personal/Subpersonal Distinction. *Philosophy Compass*, *9*(5), 338–346. https://doi.org/10.1111/phc3.12124

Dretske, F. I. (1986). Misrepresentation. In R. Bogdan (Ed.), *Belief: form, content, and function* (pp. 17–37). Oxford: Clarendon Press.

Facchin, M. (2021). Predictive processing and anti-representationalism. *Synthese*, *199*(3), 11609–11642. https://doi.org/10.1007/s11229-021-03304-3

Favela, L. H., & Machery, E. (2023). Investigating the concept of representation in the neural and psychological sciences. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1165622

Fodor, J. A. (1992). *A theory of content and other essays*. Cambridge, Mass.: MIT Press.

Fresco, N. (2014). *Physical Computation and Cognitive Science* (Vol. 12). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-41375-9

Fresco, N., & Primiero, G. (2013). Miscomputation. *Philosophy & Technology*, *26*(3), 253–272. https://doi.org/10.1007/s13347-013-0112-0

Garson, J. (2013). The Functional Sense of Mechanism. *Philosophy of Science*, *80*(3), 317–333. https://doi.org/10.1086/671173

Garzón, F. C. (2008). Towards a General Theory of Antirepresentationalism. *The British Journal for the Philosophy of Science*, *59*(3), 259–292. https://doi.org/10.1093/bjps/axl007

Gładziejewski, P., & Miłkowski, M. (2017). Structural representations: causally relevant and

different from detectors. *Biology & Philosophy*, *32*(3), 337–355. https://doi.org/10.1007/s10539-017-9562-6

Godfrey-Smith, P. (1996). *Complexity and the function of mind in nature*. Cambridge; New York: Cambridge University Press.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, *42*, 335–346.

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

Kohár, M. (2023). *Neural Machines: A Defense of Non-Representationalism in Cognitive Neuroscience* (Vol. 22). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-26746-8

Lau, H. C., & Passingham, R. E. (2007). Unconscious Activation of the Cognitive Control System in the Human Prefrontal Cortex. *Journal of Neuroscience*, *27*(21), 5805–5811. https://doi.org/10.1523/JNEUROSCI.4335-06.2007

Lee, J. (2019). Structural representation and the two problems of content. *Mind & Language*, *34*(5), 606–626. https://doi.org/10/gfkmb5

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about Mechanisms. *Philosophy of Science*, *67*(1), 1–25. https://doi.org/10.1086/392759

McGeer, V. L. (1992). The problem of error: A surd spot in rational intentionalism. *Philosophia*, *21*(3–4), 295–309. https://doi.org/10.1007/BF02380824

Miłkowski, M. (2013). *Explaining the Computational Mind*. Cambridge, Mass.: MIT Press.

Miłkowski, M. (2015). Satisfaction conditions in anticipatory mechanisms. *Biology & Philosophy*, *30*(5), 709–728. https://doi.org/10.1007/s10539-015-9481-3

Miłkowski, M. (2018). Objections to Computationalism: A Survey. *Roczniki Filozoficzne*, *66*(3), 57–75. http://dx.doi.org/10.18290/rf.2018.66.3-3

Molder, B. (2010). *Mind ascribed an elaboration and defence of interpretivism*. Amsterdam, the Netherlands; Philadelphia: John Benjamins Pub. Co. http://site.ebrary.com/id/10408503

Nanay, B. (2013). *Between perception and action*. Oxford: Oxford University Press.

Orlandi, N. (2014). *The innocent eye: why vision is not a cognitive process*. New York: Oxford University Press.

Piccinini, G. (2015). *Physical Computation: A Mechanistic Account*. Oxford: Oxford University Press.

Piccinini, G. (2020). *Neurocognitive Mechanisms: explaining biological cognition*. Oxford: Oxford University Press.

Raja, V. (2018). A Theory of Resonance: Towards an Ecological Cognitive Architecture. *Minds and Machines*, *28*(1), 29–51. https://doi.org/10.1007/s11023-017-9431-8

Ramsey, W. M. (2023). The Hard Problem of Content is Neither. *Review of Philosophy and Psychology*. https://doi.org/10.1007/s13164-023-00714-9

Rupert, R. D. (2011). Embodiment, Consciousness, and the Massively Representational Mind. *Philosophical Topics*, *39*(1), 99–120. https://doi.org/10.5840/philtopics201139116

Rupert, R. D. (2023). *Naturalism Meets the Personal Level: How Mixed Modelling Flattens the Mind*. https://philarchive.org/rec/RUPNMT. Accessed 30 September 2023

Ryder, D. (2004). SINBAD Neurosemantics: A Theory of Mental Representation. *Mind and Language*, *19*(2), 211–240. https://doi.org/10.1111/j.1468-0017.2004.00255.x

Shea, N. (2018). *Representation in cognitive science*. New York, NY: Oxford University Press.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129–138. https://doi.org/10.1037/h0042769

Thagard, P. (2000). *Coherence in Thought and Action*. Cambridge, Mass.: MIT Press.

Thomson, E., & Piccinini, G. (2018). Neural Representations Observed. *Minds and Machines*, 1–45. https://doi.org/10.1007/s11023-018-9459-4

van Gaal, S., Lamme, V. A. F., Fahrenfort, J. J., & Ridderinkhof, K. R. (2011). Dissociable brain mechanisms underlying the conscious and unconscious control of behavior. *Journal of Cognitive Neuroscience*, *23*(1), 91–105. https://doi.org/10.1162/jocn.2010.21431

Van Gelder, T. (1995). What Might Cognition Be, If Not Computation?: *Journal of Philosophy*, *92*(7), 345–381. https://doi.org/10.2307/2941061

Weichold, M., & Rucińska, Z. (2022). Pretense as alternative sense-making: a praxeological enactivist account. *Phenomenology and the Cognitive Sciences*, *21*(5), 1131–1156. https://doi.org/10.1007/s11097-021-09770-x

Yousif, S. R. (2022). Redundancy and Reducibility in the Formats of Spatial Representations. *Perspectives on Psychological Science*, *17*(6), 1778–1793. https://doi.org/10.1177/17456916221077115

Zawidzki, T. (2013). *Mindshaping: a new framework for understanding human social cognition*. Cambridge MA: MIT Press.

Zeppi, A., & Blokpoel, M. (2017). Mindshaping the world can make mindreading tractable: Bridging the gap between philosophy and computational complexity analysis. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society. Computational Foundations of Cognition* (pp. 1418–1423). Presented at the 39th Annual Meeting of the Cognitive Science Society. Computational Foundations of Cognition, Austin, TX: Cognitive Science Society.