Revista Română de Filosofie Analitică

Romanian Journal of Analytic Philosophy

Volumul XVII, Nr. 2, 2023

EDITURA UNIVERSITĂȚII DIN BUCUREȘTI

BUCHAREST UNIVERSITY PRESS

2026

Revistă semestrială publicată de Societatea Română de Filosofie Analitică,
cu sprijinul Facultății de Filosofie, Universitatea din București
Revue semestrielle publiée par le Société Roumaine de Philosophie Analytique,
soutenue par la Faculté de Philosophie, L'Université de Bucarest
Biannual journal published by the Romanian Society for Analytic Philosophy,
supported by the Faculty of Philosophy, University of Bucharest

Revista Română de Filosofie Analitică
este fosta Revistă de Filosofie Analitică (2007-2011)

www.srfa.ro

Ediția on-line: http://www.srfa.ro/rrfa/

REDACȚIA / SECRÉTARIAT / OFFICE
Universitatea din București, Facultatea de Filosofie
Splaiul Independenței nr. 204, București, 060024
E-mail: redactia@srfa.ro

ISSN (ediția electronică): 1843-9969
ISSN (ediția tipărită): 1844-2218

EDITURA UNIVERSITĂȚII DIN BUCUREȘTI

BUCHAREST UNIVERSITY PRESS

Bd. Mihail Kogălniceanu, nr. 36-46,
Cămin A (curtea Facultății de Drept),
Corp A, Intrarea A, etaj 1-2, Sector 5
050107, București – ROMÂNIA
Tel. + (4) 0726 390 815
E-mail: editura.unibuc@gmail.com
www.editura-unibuc.ro

LIBRĂRIA EUB-BUP
(Facultatea de Sociologie și Asistență Socială)
Bd. Schitu Măgureanu nr. 9, sector 2
010181 Bucureşti – ROMÂNIA
Tel. +40 213053703

TIPOGRAFIA EUB-BUP
(Complexul LEU)
Bd. Iuliu Maniu nr. 1-3,
061071 București – ROMÂNIA
Tel.: +40 799210566

DTP / COPERTĂ
EUB-BUP

Revista Română de Filosofie Analitică
Romanian Journal of Analytic Philosophy

Vol. XVII **Iulie-Decembrie 2023** **Nr. 2**

CUPRINS
CONTENTS

www.srfa.ro/rrfa/

Revista Română de Filosofie Analitică,
Romanian Journal of Analytic Philosophy
Vol. XVII, Iulie-Decembrie 2023, Nr. 2, p. 7-31
ISSN (ediția electronică): 1843-9969,
ISSN (ediția tipărită): 1844-2218
DOI: 10.62229/rrfaxvii-2/1

EMBODIMENT AND EMPATHY TRAINING

ALEXANDRU PETRIŞOR[1]

Abstract: The "Empathy Machine" hypothesis posits that virtual reality (VR) environments can significantly increase empathy more effectively than any other media. This has been widely debated, empirically tested, and proposed as a non-invasive alternative to traditional forms of moral enhancement. In this article I argue that: 1) VR has the potential to enhance empathy at least as effectively as other media, such as cinematography and literature; and 2) currently, VR does not achieve this potential. This challenges the notion that VR is the most effective medium for empathy enhancement. Following a review the literature supporting and contesting the efficacy of VR for empathy enhancement, I provide empirical and theoretical arguments to substantiate the claim that current VR projects fail to enhance empathy, and that this is due to misconceptions about embodiment in fictional settings and about different types of immersion. I conclude by suggesting that VR experiences could be as effective as other media in enhancing empathy, but present VR applications overemphasize participatory immersion at the expense of narrative and representative immersion, thereby limiting their effectiveness. The discussion underscores the need for a balanced approach to VR design that incorporates the successful elements of other media in fostering empathy.

Keywords: empathy, virtual reality, embodiment, make-believe, moral understanding.

[1] Alexandru Petrişor is a doctoral student within the Faculty of Philosophy at the University of Bucharest.

Introduction

Let us call the "Empathy Machine" hypothesis the claim that by using virtual reality (VR) environments we can significantly increase people's empathy much more effectively than with any other media. This hypothesis has received much discussion in the literature, being the test target for many studies and being seen as a non-invasive alternative to other forms of moral enhancement (Zahiu et al., 2023). In this article I will argue: 1) that VR could enhance empathy at least as well as other media (cinematography, literature, etc.), and 2) currently this is not the case. (1) establishes a lower limit for well-constructed VR experiences that it shares with other media, while leaving the door open for it to be even better. (2) serves to deflate the premature optimism of the 'Empathy Machine' hypothesis. In the first part I will elucidate each conjunct of my thesis. The first one is the target of many articles in the literature (Zahiu et al, 2023; Bloom, 2017b; Lara & Rueda, 2021; Rueda & Lara, 2020; Sora-Domenjó, 2022; Martingano et al., 2021), although the formulation can differ depending on what one takes to be the case. If I think that VR is an empathy machine, then I will argue for a stronger variant of (1): that VR is better than other media at enhancing empathy. Alternatively, if I believe that VR is an inadequate medium for enhancing empathy, then I will argue that it isn't nearly as good for said purpose as other forms of media. There may be other possibilities but the discussion will reveal that (1) is the best formulation of this thesis because it keeps VR in line with other media. Regarding (2), my argument will be based both on empirical findings and on armchair theorizing about how current VR projects miss the mark when trying to enhance empathy. My speculative suggestion is that this failure stems from misconceptions of both embodiment in fictional settings and immersion. This will work as a how-possibly explanation of why VR isn't the empathy machine many authors expect it to be. As such, I will try to better articulate what are the relevant factors that go into empathy enhancement via VR experiences.

The article is divided into five sections. The first one defines the terms that I have hitherto left to their intuitive meaning in order to better understand the issue at hand. Also in this section, I will comment more on the motivation behind philosophical discussions about empathy enhancement. The next section deals with some familiar accounts from the literature on VR and empathy. The third section contains the arguments supporting my thesis, dealing with (2) first and (1) subsequently. As we will see, the reversal is well-motivated. In the fourth section, I will reject several counterarguments to my thesis. The fifth section concludes the article.

1. Defining VR and empathy

1.1. Virtual Reality (VR)

Since Virtual Reality is so popular both in recent academic literature and because of devices that make their way into the public's hands, I need to clarify what I mean by VR. We can describe VR as a medium that uses computer modeling to simulate 3D environments with which the player/agent can interact. I will focus on HMD VR devices (head-mounted displays), but the discussion isn't limited to this type of technology. This step just lets us more easily refer to current widely available technology. A plethora of VR studies use such an interface (Martingano et al., 2021), but an HMD does not differ fundamentally (although the effect on the feeling of being there in the virtual world is magnified by its use) from a simple computer screen – and this may matter when taking into account how interactive movies function. Nonetheless, my arguments are not limited to HMDs.

One might wonder what makes a VR experience fundamentally different from a video game played on a regular display. Here, Chalmers' observations can help: "What is distinctive about VR is that its virtual worlds are immersive. Instead of showing you a two-dimensional screen,

VR immerses you in a three-dimensional world you can see and hear as if you existed in it" (Chalmers, 2022, p. xii). Echoing our definition, Chalmers describes VR environments (there is some difficulty in extending this definition to VR experiences, as I will later suggest) as immersive, interactive, and computer-generated. So, immersion is a key component of VR experiences, inasmuch as they presuppose VR environments. From Chalmers' quote, we can gather that the relevant sense of immersion is perceptual: the fact that you can see mountains and hear the gust of wind is what makes it an immersive environment. As we will see, this is just one type of immersion to consider when talking about experiences in VR environments.

Presumably, by using computer modeling you could, in an intuitive sense, be put in the shoes of any other being with similar enough perceptual systems. For example, by the use of VR, you could be in the shoes of a Syrian refugee (Arora & Milk, 2015), or a black man in some unspecified US city (Cogburn et al., 2018). It doesn't seem to be the case that you could be put in the shoes of a bat if that would mean getting to know what it is like to be a bat (Nagel, 1974). Being in someone's shoes is not that clear; in the third section, I'll advance several ways to cash it out.

1.2. Empathy

Many definitions of empathy are available. In different corners of the literature, empathy gets assigned different roles/importance: as a way of getting to know what somebody feels, as emotional contagion, as getting to know *what it is like* for someone to feel something, etc. (Gallagher, 2020; Zahavi, 2014). It would help to use a definition that doesn't presuppose too many specific choices about the role and importance of empathy. This is why I follow Zaki and Ochsner, taking empathy to be "the ability and tendency to share and understand others' internal states" (Zaki & Ochsner, 2016, p. 871). So construed, empathy can be described as having three components:

1) experience sharing = the tendency of perceivers to take on the sensorimotor, visceral, and affective states of the targets;

2) mentalizing = perceivers' explicit reasoning about targets' internal states using lay "theories" about how situations produce internal states;

3) prosocial motivation = through which individuals who share and understand targets' states often are compelled to help those targets. (Zaki & Ochsner, 2016, pp. 871-872)

The definition of empathy describes empathy as both ability and tendency: (1) is a tendency, whereas component (2) is an ability. This implies empathy is not a "success term" (Walton, 2015, p. 2). You may empathize with X but wrongly attribute to her the mental state Y, instead of Z. Presumably, your failure to properly empathize would stem from your lack of ability in empathizing. Maybe you need more training: the empathy trainer may prescribe reading more novels or watching more movies. (This raises questions regarding the individuation of mental states and attitude ascriptions, respectively.) Perhaps terms for abilities can sometimes be success terms too – if you fail to initiate the process at all; no representation means neither success nor failure. As you grow more competent in your empathizing skill, you gain the ability to attribute the correct mental state in cases where previously you would have been silent. Even so, the point of characterizing empathy in dispositional terms is that it can be improved. Given empirical data supporting this hypothesis (Persson & Savulescu, 2018), I bracket worries about the most intuitive way of characterizing empathy and accept the given definition as a working one.

Although (2) explicitly mentions lay "theories", we should not take this to mean a commitment to theory of mind (ToM) accounts. I use "mentalizing" to refer to whatever mental processes that help us attribute mental states to others, in ways *somewhat* linked with experience sharing and motivational aspects of empathy. Using "mentalizing" this way sets aside cases of experts at attributing mental states to others who lack the

motivation to help them and/or appropriate affective reactions to moral violations (such as anger/disgust towards harm done to innocents). Such cases are paramount in evaluating an individual's specific moral profile, but fade when we consider moral behavior on average, in the general public.

2. The Empathy Machine

Having fixed working notions of empathy and mentalizing, I now turn to the imaginative aspect of empathy, the focus of debates surrounding empathy enhancement. Intuitively, some people simply fail at imagining how others feel and therefore fail to understand their mental states. This, in turn, can lead to disastrous consequences. Zahiu et al. put the problem this way when they distinguish between bounded and reflective empathy (2023, p. 4), which correspond to experience sharing and mentalizing, respectively. Bounded empathy is an evolutionary strategy of behavior optimization that is adaptive to the challenges of a given environment. It is triggered automatically and spontaneously biased (see Bloom, for more on the biased aspects of empathy). Reflective empathy is described as being "supervised by voluntary acts of imagination and deliberation" (Zahiu et al., 2023, p. 5) and thus under cognitive control. Is there room for improvement on the experience sharing/bounded empathy side? Maybe, but it is hard to see how, given its evolutionary design and biases (Bloom, 2017a). Since humans have a lot of cognitive resources that can be leveraged, and since training can help improve many other cognitive-based abilities (such as playing chess), we might have better chances at improving our mentalizing skills, rather than our experience sharing ones:

> "What VR technology can do is to scaffold our imaginative powers. VR technologies are uniquely equipped to attach vivid sensory representations to experiences and interactions. It can bring certain experiences in the here and now, instead of letting individuals rely on imagined hypotheticals. This, in turn, has the potential to enlarge our moral understanding." (Zahiu et al., 2023, p. 6).

Let us now use this passage as representative of empathy enhancement with VR, and refer to it in a short-handed manner as "(EM)". It is not clear if (EM) targets experience sharing or mentalizing. On the one hand, it seems to be explicitly about mentalizing, since it involves imagination as used for understanding the mental states of others. Presumably, this is to be done by imagining oneself in the shoes of the target. On the other hand, (EM) claims that, by making people subject to vivid sensory representations of certain experiences/interactions, we avoid the need to imagine said circumstances. On the supposition that mentalizing requires imagining, or at least some degree of deliberation, it seems that VR eschews both. If so, the person connected to the empathy machine does not mentalize anything, but instead shares an experience with some target. This does not mean that (EM) is defective, since the relevant VR experiences might not involve representations so vivid as to forestall the need for any mentalizing or deliberation. But this is an empirical and technological question, one that can be probed by looking at studies and checking whether VR training increases empathy (specifically via which component), and how data might vary with newer generation VR.

This is very important for my thesis. Since empathy is described as having three components, it is at least conceivable that each can operate in the absence of another. This isn't entirely true if one looks at the neural correlates of said components, even though the neural circuitry differs significantly with respect to the first two (Zaki & Ochsner, 2016). One might have a pronounced tendency to share the experiences of others, so she will naturally pick up the mental states (emotions/moods) of other people when they are in her proximity. It is at least conceivable that such a person could perform poorly on mentalizing tasks, especially ones that require her to empathize with people not in her proximity and differing in some non-essential characteristics enough so as to not be perceived as a member of an in-group. Evidence from studies also suggests that this is not just a possibility, but is the actual case for some people (Zahiu et al.,

2023). Likewise, one might be a master mindreader even though she is not able to easily engage in experience sharing. Evidence suggests this is the case for some (Rijnders et al., 2021). Since these two can come apart to some extent, we can raise the question of what exactly it is that VR experiences train.

Compare: literature, too, tends to increase mentalizing skills, as opposed to increasing the disposition to share an experience (Martingano et al, 2021). What (if anything) makes VR (but not old-fashioned reading) an Empathy Machine? Bounded empathy is prone to biases, corrigible if people rely more on reflection. But if VR does not supply us with the right increase in the right parameter (e.g., reflective empathy) then its label as an Empathy Machine seems unwarranted. Worries also linger about the ethics of using VR experiences for increasing empathy:

> "The pleasures of toxic embodiment offered by witnessing racial suffering in VR extend these precarious conditions of life. Rather than trying to automate compassion for those who suffer as a result of racialized violence, indifference, and hyperautomation, we need forms of critique that show us how emotions like empathy and compassion have *alibi*-ed untenable material conditions of labor for racialized and gendered people long before VR claimed them." (Nakamura, 2020, p. 61).

Nakamura targets VR experiences that aim to put a player in the shoes of a minority. Not all VR experiences aim to do this and there is no requirement that they should, even though it seems to be an easy choice to make given the way the world is right now.

These discussions lay a strong emphasis on embodiment and on identifying the virtual agent with the person undergoing the VR experience. You are supposed to be in the shoes of a refugee and that means that you are supposed to (fictionally) have such-and-such characteristics. This is facilitated by vivid representations you are supplied with. But, we might ask, what game of make-believe is played when engaging with such fictional worlds? The lack of this kind of analysis is felt on both sides. Presumably, Nakamura would not have any

problem with literary works of marginalized people describing their first-person experience. Is this to be explained by virtue of the literary medium itself, or by features of the types of games of make-believe that one engages in when reading said works? On the other hand, is empathy enhanced by vivid representations, or in virtue of the game of make-believe that VR experiences authorize? These questions shift focus away from the surface characteristics of the VR medium to the way in which users participates in it.

An analogy might help: the way the current discussion goes is akin to saying that paintings done in a realistic style are better at making one understand the world of a painting than paintings done in an impressionistic style. This ignores the onlooker's role in engaging in games of make-believe with said paintings, and as a result one might misidentify the parameter X (let us say proficiency in understanding paintings for the sake of the example) responsible for the increase in skill Y (understanding the fictional world of a painting) with some characteristic related to Z (painting style), that nonetheless is not essential to X or Y.

(EM) is also a thesis about a type of understanding: moral understanding. Empathizing with someone not only helps get a sense of how that person feels but also 1. functions as a motivator for doing something about it (to remove the cause of harm) and 2. gives rise to first-hand propositional knowledge (or know-how if you are an anti-intellectualist when it comes to skills and abilities) that is then used in moral and action-guiding inferences, knowledge that is then applicable in multiple structurally similar cases. You gain moral understanding when you have "a grasp of the connections between moral reasons and moral conclusions" (Hills, 2020). As construed here and as I believe that (EM) is intended, VR helps develop moral understanding because it targets the cognitive aspect of empathy. This is what makes possible the multiple applicability of empathy-triggered inferences (that the same

harm is being done in cases of torture even though the subjects have different skin colors).

Remember that we assume, in line with Zahiu et al., that bounded empathy is biased and unlikely to be corrected. What makes one grasp moral knowledge should enable assenting to the moral structural similarity of cases X and Y, differing only in characteristics such as the skin color of the victim. It also should also allow one to generalize their findings: after being exposed to the harsh life of a refugee, one should draw universally quantified conclusions, that all such acts are bad, etc. One could then apply this new-gained knowledge and understanding so as to deal with new cases. In contrast, bounded empathy limits itself to one individual or one group. This is why we should focus on the measurable increase of cognitive empathy (mentalizing). As we will see, there are strong reasons to doubt that current VR experiences are able to fulfill (EM). And this is to be explained, I contend, by employing Walton's (1990) influential analysis of fiction and representational works of art.

Empirical data analysis does not support (EM) unqualifiedly. One recent meta-analysis has found that VR interventions improved bounded empathy/experience sharing, not reflective empathy/mentalizing (Martingano et al., 2021). This meta-analysis surveyed 43 studies with a total of 5,644 participants, so we can take it at least as a worrying result for (EM). One point worth mentioning is that the authors also included presumably less immersive interactive movies. This could be seen as a limitation. The authors, however, found that "more immersive delivery devices that used head mounter displays did not have a significantly larger effect on empathy than non-immersive delivery devices that ran on a normal computer desktop or headphones" (Martingano et al., 2021, p.13). It seems unlikely to say that VR could be an empathy machine and yet fail to increase the ability to mentalize. It seems even less likely that VR could be such a machine in virtue of its immersive embodiment and vivid representations and yet fail to elicit better results than interactive movies which have a lower degree of bodily immersion.

I aim to argue that VR experiences could be just as good as other media for increasing empathy (focusing on the mentalizing aspect) but currently, fail to realize that potential. This meta-analysis supports the second conjunct. I now turn to elucidating the important characteristics of the games of make-believe that VR experiences authorize in order to support the first conjunct.

3. Embodiment and immersion

VR experiences differ from those in other relevant media given immersion and the feeling of embodiment VR procures (Chalmers, 2022). Immersion and embodiment *cannot* be used interchangeably, a lesson sometimes ignored (e.g., Sora-Domenjo, 2022, p. 6, which contrasts immersive experiences with 2D movies). I hold that most current VR experiences lack a focus on narrative immersion[2], as opposed to embodiment (perceptual/spatial immersion) and a lack of agency given to the participant in such experiences. In Walton's (1990) terms, a player/participant who engages in a VR-authorized game of make-believe may nonetheless lack representational capacities needed to explain their actions and *de se* imaginings. In the following subsections, I will argue that embodiment does not guarantee immersion and I will illustrate by means of VR experience samples the ways in which they fail to prescribe the adequate *de se* imaginings that will engage mentalizing.

3.1. Four dimensions of immersion

Being immersed in something that can be described as a state in which one 1) is attentive to the activity in which one is immersed, 2) fails to pay

[2] As will be detailed below, a VR experience is narratively immersive if and only if the subject's experience in VR has a coherent narrative structure and flow. To illustrate with a counterexample: if your favorite historical movie abruptly interposed the story of Spider-Man halfway through, it would lack narrative immersion.

attention to her surroundings if these don't pertain to the activity she is immersed in. I take immersion as a property of one's engagement in an activity. This should not be controversial, as one could define derivative ways in which they are immersed in, for example, a culture by reference to activities pertaining to said culture. Being immersed in a fictional world would then mean that the player pays attention to the game of make-believe she takes part in.

I treat two VR experiences as a sample representative for the whole set of currently available VR experiences which could be used to enhance empathy. One is *Clouds over Sidra*, an interactive movie in which one is presented with the life of Syrian refugees, and the other one is *1000 Cut Journey*, in which one is put in the shoes of a black man as he experiences instances of racism. I choose these two thanks to their popularity in the academic literature and beyond. *Clouds of Sidra* is an interactive movie, but that should not detract from the points I will make, since it functions almost exactly like *1000 Cut Journey*, with the exception that you cannot throw a basketball at any point in the experience. Should that make for a qualitative difference between the two? I see little reason for this claim. *Clouds of Sidra* asks the participant to follow the speaker, and this is done by turning your head around if you are using an HMD. So, there is still a feeling of embodiment/presence in the experience, even though it isn't as complex as in *1000 Cut Journey*. Both experiences allow the user to participate by moving their head in inspecting the virtual environment. In both cases, the characters with which they interact are recordings of real actors playing their designated roles. As such, there isn't much interaction between the player and the non-playable characters, but there is a high representational fidelity of said recordings as props. Due to the likeness that they bear to real experiences, they can easily prompt imaginings regarding their facial expressions or body language. Presumably, this was meant to avoid the uncanny valley effect, which might have influenced empathic responses. Nonetheless, representational accuracy (of people, of interactions) trades off against the level of participation the player exhibits.

The trade-off matters. If the VR experience supplies me with the highest-fidelity sensory information I might feel immersed in the sense that I find it difficult to come to believe that what I am experiencing is not the real world but just computer-supplied images. But if I cannot engage properly with said world – if I cannot walk over to the basketball course and throw a basketball, if I cannot talk to other characters within the world – then I will find it less difficult to come to believe that I am only in a VR experience[3]. Immersion should, then, not be left to intuitions unchecked by conceptual analysis. Balcerak Jackson & Balcerak Jackson aptly note: "There is nothing wrong with such metaphorical characterizations, of course, but their utility is limited when it comes to thinking systematically about VR and our engagement with it" (2023, p. 19).

To illustrate, Zahiu et al. seem at first blush to equate immersion with the feeling of embodiment (2023, p. 3). Likewise, Zaki contrasts VR with other "non-immersive" forms of media, like literature (2014). Nakamura (2020) views VR as allowing voyeuristic engagement; I take this to mean that it does not fully immerse oneself in the experience of a marginalized individual. Bloom (2017b) also emphasizes this lack of full immersion and its replacement with a voyeuristic type of behavior. Such stances err in taking immersion to consist solely in the feeling of embodiment that a participant has when engaging in VR experiences.

(Balcerak Jackson & Balcerak Jackson, 2024) distinguish four different types of immersion: 1) representative immersion; 2) participatory immersion; 3) affective immersion; and 4) narrative immersion. A VR experience is representationally immersive iff it involves a rich and coherent network of perceptual or affordance representational states. VR experiences with photorealistic graphics should rank high on representational immersion. The affordance condition covers more complex mental states that should be prompted by the experience. For instance, being in a VR environment that resembles a city should generate specific representations/affordances (e.g. crossing

[3] I will leave aside differences between expert VR users and novice ones (Chalmers, 2020).

the street to get to the bakery). Balancing on a plank (in a VR environment) generates the state of believing that one is going to fall if one were to step on the side of the plank.

A VR experience is participatorily immersive iff the participant can act so as to determine what the virtual environment is like and how events unfold in it. So defined, participatory immersion differs from what Chalmers calls interactivity. *Clouds over Sidra* is still participatorily immersive because one can choose how to position themselves in the scene and where to focus their attention. This can be extended to literary works: you can choose how to imagine some details from how a room is described (as they are not determinate enough to mandate one single interpretation), therefore you participate in the creation of the fictional world that you are engaging with.

A VR experience is affect immersive iff one emotionally or affectively participates in the virtual experience. This is a stronger condition than merely requiring that a user be psychologically participating in the virtual experience. Imagine disinterestedly playing a VR game in which you have to jump platform to collect coins. Little if any emotion is felt whilst doing the boring task. Nonetheless, one still believes that virtually, there is a platform with ten coins over there, and that after collecting said coins one can proceed to the next level.

Lastly, a VR experience is narratively immersive iff the subject's experience in VR has a coherent narrative structure and flow. This is hard to make precise given the many different types of narrative structures out there. The important point is that this is a type of immersion that differs from the previous three. As we will see, narrative immersion is what many VR experiences ignore.

3.2. The narrow focus on embodiment

My argument for why many current VR experiences fail to deliver improvements in mentalizing is the following:

1) Current VR experiences are developed with the assumption that embodiment guarantees immersion in the relevant sense.

2) Current VR experiences are developed with the assumption that immersion in the relevant sense guarantees improvement in mentalizing.

3) Embodiment does not guarantee immersion in the relevant sense.

4) Therefore, current VR experiences do not improve mentalizing (thus, explaining the empirical data).

Obviously, this is a schematic version, as there isn't such a thing as immersion without a work in which a participant is immersed, and that work has to have certain characteristics in order to target mentalizing. Appropriate kinds of immersion also need to be specified. Nonetheless, this formulation is enough to get the point across. Pending filling in the details, the argument seems valid. I will primarily consider the set of immersive experiences that improve mentalizing (so I eliminate player experiences of immersion in the role of a super-soldier shooting aliens). "Immersion in the relevant sense" sets aside works that only give a representative sense of immersion, which is unlikely to improve mentalizing. Similarly, works that only target affective immersion might improve bounded empathy but not reflective empathy. Nobody would hold that a VR experience with photo-realistic graphics, but nothing to do with it but walk around, is relevant to developing empathy. (1) and (2) come from my reading of the literature and I already presented evidence for their plausibility. I will therefore argue for the truth of (3) in order to show the argument is sound.

Friends and foes of VR as a medium for empathy enhancement presuppose that immersion has to do with the sense of embodiment one experiences when putting on an HMD. This is only partially correct. It is more immersive to have a first-person perspective in a VR environment than engaging with it via a detached computer screen, all other things being equal. However, the *ceteris paribus* clause obscures the complex interaction between different aspects of a VR experience that help foster the sense of immersion. In our sample of VR experiences, even when there is a strong sense of embodiment, either the experience elicits emotional

responses for other reasons (the sad music present in *Clouds over Sidra* and the performance of actors in both this experience and *1000 Cuts Journey*) or the experience fails to elicit emotional responses because it neglects other relevant factors; for example, how *1000 Cuts Journey* undermines its own goals by failure to build narrative coherency and not letting the participant engage in certain types of mental actions that would foster mentalizing and a sense of representative immersion.

A VR experience isn't identical to a VR environment. The former designates the authorized games of make-believe to be played with the VR environment, whereas the other only points to a set of props that are to be integrated into that VR experience. Moreover, the VR environment can include elements that do not feature in the VR experience. Does representative immersion enhance empathy? Surprisingly, this is often taken for granted. Your empathizing with the refugees *in Clouds over Sidra* owes a lot to the sad music playing in the background and the authentic depictions of actual refugees. Being in the shoes of a refugee doesn't mean that you hear sad violin music throughout. Sad music helps form affective immersion but it sacrifices representative immersion.

It is not fictional that the refugees (including the player) can hear the sad music, therefore the music isn't part of the VR experience, even though it is part of the VR environment. Hearing sad music when witnessing people gathered around a table is more likely to elicit imaginings about funerals than hearing upbeat happy music when presented with the exact same scene. This one aspect perfectly illustrates that the VR environment helps frame the VR experience by setting up an intended interpretation (the authorized game) of what is to be fictionally felt and believed.

But if our goal is to enhance cognitive empathy, having the right music playing in the background is similar to solving a test and having the answers written on the back of the page. It leaves little room for the participant to figure out for herself what those fictional people are feeling. In this sense, it comes across as second-hand knowledge: sad music works as a testimonial basis for forming the belief that what the characters are

experiencing is sadness. It gives you a reason to believe something, as opposed to merely prompting you to inquire into what is fictionally true.

Do real people in VR enhance empathy and moral understanding? Or, rather, recordings of real people, as opposed to digital characters that are coded to respond to what the participant does. Their facial expressions help trigger emotions in the participant, but once again little room is left for mentalizing when you can directly see sadness on display. Add to this the disembodied narrator's voice present throughout, and one cannot help but be reminded that one is experiencing a performance put on display for her/him.

In VR experiences such as our two samples, disembodied narrator voices function as an egregious case of an "aside to the audience" (Walton, 1990). This breaks immersion (all kinds) by making the participant aware that one is playing a game of make-believe with a certain work. In the words of Chasid, one's experience with the fictional world becomes "doxastically mediated" (2023). The participant becomes aware of the rules by which she is playing the game of make-believe which is her VR experience. To illustrate, at the beginning of *Clouds over Sidra*, one is directly addressed by Sidra, the girl refugee who is the main character in the work. Then the perspective shifts to different scenes showing a day in Sidra's life, and the participant can only hear her disembodied voice whilst witnessing said scenes. If, in the beginning, one could have a sense of deep immersion (narrative but not only) generated by implicitly understanding what is presented as fictionally true and, therefore, to be imagined, later the rules of what is to be imagined become opaque and doxastic mediation sets in. Is it fictionally true that Sidra walks you through these various scenes, or are you presented with what you are supposed to imagine whilst talking to her?

Other factors may help you stay on track (the narrative component of immersion is developed through Sidra's narration) and affectively immersed in the experience, but the fact that you embody a refugee certainly is of no help here. A character could be facing you and talking to you, the audience member, even if you are watching it from a movie

screen, not an HMD. Psychological participation is still high even absent a sense of embodiment. Embodiment does not guarantee immersion: there is an acute need for other factors as well (coherency of narrative, promptings for affective responses and inquiries into morally relevant factors).

Thus far, I focused on *Clouds over Sidra*. But *1000 Cuts Journey* makes more serious mistakes. Once again, you are presented with a disembodied narrator voice that generates doxastic mediation of what is to be imagined. The feeling of embodiment is more acute: at one point you are in front of a mirror, at another you can throw a ball and some cubes. All this is insufficient for the relevant sense of immersion that enhances mentalizing, for embodiment could have equally been experienced this way in a game about jumping on platforms to collect coins. In terms of narrative coherency, the experience is lacking.

In the end, you are given a moral lesson from the narrator's voice saying that this is the experience for many people in the world and that they hope you leave the VR experience being more open to evidence of racism. Literary works do tend to display sentences such as "X is bad", but there is a way of doing it that does not come across as moralizing or as giving away the answer to the test question (as second-hand knowledge). There are many ways to avoid this, but they essentially involve *propagating*, not *transmitting* knowledge. Knowledge is transmitted (by testimony) when "the speaker's knowledge that *p*, expressed through her testimony, is an epistemic ground of the learner's knowledge that *p*" (Hills, 2020, p. 401). In contrast, propagation of knowledge takes place when "the speaker's knowledge that *p*, expressed through her testimony, is a causal factor and an epistemic influence on the learner's gaining knowledge that *p*, but not an epistemic ground of it" (Hills, 2020, *ibidem*). It is important to distinguish transmission from propagation, as the first one yields second-hand knowledge, whereas the latter, by prompting the learner to herself go through the process of gaining the knowledge that *p*, yields first-hand knowledge.

The participant in a game of make-believe is a reflexive prop and generates fictional truths about himself (*de se* imaginings). It is fictionally true in the game of make-believe that I play with *1000 Cuts Journey* that I am a black man and live through such and such experiences. Moreover, I imagine this *from the inside*. I do not merely imagine that I see a police officer, I am imagining *seeing* the police officer. Embodiment is an important aspect of immersion because it prompts imaginings *from the inside*. You get to imagine *throwing* a basketball, not just that you throw it. But that is not enough. Disembodied narrators transmitting knowledge rob the participant from thinking through/inferring/inquiring about the situation they are in, despite feeling at all times embodied. Lack of representative immersion undermines enhancement in mentalizing.

The way moral knowledge is transmitted in *1000 Cuts Journey* also emphasizes the experimental setting in which these experiences are had: if immersion is what you are after, why would you emphasize that what the participant has just witnessed is a VR experience done at X university under the supervision of Y professors and researchers? I take this to suggest a lack of focus on the narrative the participant is supposed to experience. The participant in the *1000 Cuts Journey* authorized game of make-believe is faced in 10 minutes with three different episodes from the life of supposedly the same man. No other character makes more than one single appearance. What ties together these three episodes is the narrator giving prescriptions to imagine these as episodes of the same man. There is barely enough time to think about what is happening before you are given the imperative to combat racism. This is clumsy storytelling, and it shows in the fact that participants are not left with improvements in mentalizing skills after undergoing the experience (as per section 2). The feeling of embodiment is present throughout, but because of a lack of narrative immersion and representative immersion when it comes to the sorts of mental actions that a participant can fictionally go through (inquiring whether how skin color could ever warrant bad behavior from others), empathy enhancement and moral understanding are undermined.

We now have good reason to conclude (3) is true. Mere embodiment does not guarantee the intended sense of immersion that would foster improvements in mentalizing. There is little reason to believe that VR experiences are much better at this task than any other media – unless one wants to argue that people are more prone to undergoing VR experiences than anything else, which is an empirical question, and intuitively the answer is negative. The truth of (3) fits well with the data presented in section 2 and it also has the merit of identifying what relevant factors go into making a VR experience that could improve mentalizing. Only those factors are precisely the ones that VR as a medium shares with other media: narrative coherency and the ability to prompt *de se* imaginings that would foster moral understanding.

4. Replies and objections

I have assumed that developments in mentalizing depend on a certain level of narrative and representative immersion. Is the assumption unwarranted? Mentalizing has been much discussed in child development studies and social cognition, and might even be a misnomer (some ToM critics would charge that we rarely mentalize [Gallagher, 2020]). I stay neutral on how to best construe ToM, and on which account of social cognition is preferable. The foregoing does not presuppose a ToM-based approach: 1) direct social perception is not ruled out, being available for some sorts of VR experiences (such as those with live-action characters); 2) my view explicitly deals with narratives and particulars, not just with generalizations and reasoning about instances based only on inferential rules; 3) any emphasis on inferences, in order to cash out a moral understanding thesis, goes well beyond ToM and social interaction theories, since moral problems go beyond simply knowing how to interact with others. As such, my account is pluralistic about social cognition. Whether you end by call narrative-based inferences distinct in

kind or not from theoretic inferences does not undermine my discussion of VR empathy enhancement. I explicitly focused on representative immersion regarding the sorts of mental acts that people undergo when coming to believe that "X is bad" because moral understanding presupposes such acts and is essential to developing the mentalizing aspect of empathy. Narrative immersion was emphasized because we do not attribute mental states to people in a void, but always in a social context, relating them to their goals and moral evaluations besides many other factors. Add to this that engaging in games of make-believe and other such pretend play is a great opportunity for self-knowledge (via *de se* imaginings), and the necessary links between narrative coherency and developments in mentalizing and moral understanding via VR experiences (or any other media for that matter) become visible.

How is a participant supposed to develop, on my account, the necessary know-how to reach a moral understanding? Presumably, one has become an expert empathizer with respect to mentalizing. VR experiences might afford an efficient know-how knowledge transfer, thanks to their specific medium characteristics. I take no stance on this issue, either intellectualist or anti-intellectualist, as I believe that both could be easily integrated into discussing the different types of immersion and how they contribute to empathy enhancement. One good argument in favor of my account, as opposed to one that emphasizes the essential know-how that comes with the agentive aspect of VR embodiment, is that mentalizing is significantly improved when one reads novels or watches movies or plays. There is no embodiment there, but still one either acquires the propositional knowledge that such-and-such people feel such-and-such emotions when they find themselves in certain situations or, alternatively put, one still gets to know how to react in certain situations based on what one has fictionally experienced.

One could say that I have not proved that mere embodiment does not guarantee immersion in the relevant sense, but instead, I only managed to prove that these samples do not push embodiment to its full

potential. Embodiment is linked with a sense of agency. Being an agent presupposes certain actions being available, in the context of the narrative, to the participant. These samples do not provide said affordances, therefore they should not be brought up when discussing the merits of VR as an empathy enhancement machine. This transcendental argument of sorts against my position fails because it leaves the notion of agency completely open to interpretation. When are we expected to feel like a proper agent in a VR environment? How big is the set of actions that should be available to the player? Should a player be able to role-play anything in a VR environment in order for that experience to be representative of the medium? VR experiences have a lot in common with video games. Both have virtual environments that the player inhabits and limits on which courses of action can be pursued. Still, there are limitations to what a game designer can allow a player do to. Often, the intended story that the designer is trying to communicate limits what actions players can perform. This is no fatal flaw of the game experience, as the limited agency isn't perceived as immersion-breaking unless it inhibits the narrative coherency or flow of the game. For instance, if you are playing a game in which you are the hero that is supposed to save the world and the end is near, immersion could be broken by the fact that due to a lack of an in-game timer, you can spend years (fictionally) doing chores and other fun activities, whilst ignoring the main quest. But the fact that you cannot build cars in the same game does not detract, *ceteris paribus*, from narrative immersion, although it could be counted as a downside of representative immersion. The transcendental argument simply presupposes that maximizing representative and participatory immersion is all there is to the agentive aspect of VR experiences, and that aspect is key to developing empathy. That would be mistaken; for a narrow focus on representative and participatory immersion leads to an incoherent narrative, which in turn, leaves too little room for developing mentalizing skills.

Conclusion

VR experiences could, in principle, be just as good at enhancing empathy. Currently, VR has not realized that potential. The assessment relies on empirical evidence from studies on empathy enhancement in VR environments. I have provided a how-possibly explanation for the data, while also highlighting the complex interaction between the different types of immersion when one engages in a game of make-believe with a VR environment. As I argued, VR experiences focus too much on participatory immersion through embodiment, sacrificing both representative immersion and narrative immersion. Representative immersion is undermined through a limited virtual environment that lacks promptings for mental acts that would foster moral understanding and mentalizing. Narrative immersion is undermined by an over-reliance on explicit narration and on creating a tight experimental setting which manages to only create a second-hand transmission of the relevant moral knowledge. This, in turn, leads the participant to interact in a doxastically-mediated way with their imaginative projects. Instead, I argued that the same factors that foster developments in mentalizing in other media are also at play in VR experiences: narrative coherency, affective engagement, and prompting moral inferences and inquiries. Theorizing them will be key in analyses on how future VR technology fares when compared to other media.

References

Arora, G., & Milk, C. (2015). *Clouds of sidra*. Retrieved from https://www.with.in/watch/clouds-over-sidra/

Balcerak Jackson, Magdalena & Balcerak Jackson, Brendan (2024). Immersive Experience and Virtual Reality. *Philosophy and Technology* 37 (1):1-24

Bloom, P. (2017a). *Against empathy: The case for rational compassion.* Random House.

Bloom, P. (2017b, February 3). It's Ridiculous to Use Virtual Reality to Empathize with Refugees. *The Atlantic*. Retrieved April 14, 2022, from https://www.theatlantic.com/technology/archive/2017/02/virtual-reality-wont-make-you-more-empathetic/515511/.

Chalmers, D. J. (2022). *Reality+: Virtual Worlds and the Problems of Philosophy*. Penguin UK.

Chasid, Alon (2021). Imaginative immersion, regulation, and doxastic mediation. *Synthese* 199 (3- 4): 1-43

Cogburn, C., Ogle, E., Bailenson, J., Asher, T., & Nichols, T. (2018). *1000 cut journey*. Retrieved from https://vhil.stanford.edu/1000cut/

Gallagher, S. (2020). *Action and Interaction*. Oxford, United Kingdom: Oxford University Press.

Hills, Alison (2020). Moral Testimony: Transmission Versus Propagation. *Philosophy and Phenomenological Research* 101 (2):399-414.

Lara, Francisco & Rueda, Jon. (2021). Virtual Reality Not for "Being Someone" but for "Being in Someone Else's Shoes": Avoiding Misconceptions in Empathy Enhancement. *Frontiers in Psychology*. 10.3389/fpsyg.2021.741516

Martingano, A. J., Hererra, F., & Konrath, S. (2021). Virtual Reality Improves Emotional but Not Cognitive Empathy: A Meta-Analysis. *Technology, Mind, and Behavior*, 2(1). https://doi.org/10.1037/tmb0000034

Nagel, T. (1974). What Is It Like to Be a Bat? *The Philosophical Review*, 83(4), 435–450. https://doi.org/10.2307/2183914

Nakamura, L. (2020). Feeling good about feeling bad: Virtuous virtual reality and the automation of racial empathy. *Journal of Visual Culture*, 19(1), 47-64

Persson, I., & Savulescu, J. (2018). The moral importance of reflective empathy. *Neuroethics*, 11(2), 183–193. https://doi.org/10.1007/s12152-017-9350-7

Rijnders, R. J. P., Terburg, D., Bos, P. A., Kempes, M. M., & Honk, J. van. (2021). Unzipping empathy in psychopathy: empathy and facial affect processing in psychopaths. *Neuroscience & Biobehavioral Reviews*, 131, 1116-1126. doi:10.1016/j.neubiorev.2021.10.020

Rueda, J., & Lara, F. (2020). Virtual reality and empathy enhancement: Ethical aspects. *Frontiers in Robotics and AI*, 160. https://doi.org/10.3389/frobt.2020.506984

Sora-Domenjó, Carles (2022). Disrupting the "empathy machine": The power and perils of virtual reality in addressing social issues. *Frontiers in Psychology* 13

Walton, Kendall (1990). *Memesis As Make-Believe*. Harvard University Press.

Walton, Kendall L. (2015). In *Other Shoes: Music, Metaphor, Empathy, Existence.* New York: Oxford University Press.

Zahavi, D. (2014). Empathy and Other-Directed Intentionality. *Topoi* 33 (1):129-142.

Zahiu, A., Mihailov, E., Earp, B. D., Francis, K. B., & Savulescu, J. (2023). Empathy training through virtual reality: moral enhancement with the freedom to fall? *Ethics and Information Technology*, 25(4). https://doi.org/10.1007/s10676-023-09723-9

Zaki, J. (2014). Empathy: A motivated account. *Psychological Bulletin*, 140(6), 1608. https://doi.org/10.1037/a0037679

Zaki, J., & Ochsner, K. (2016). Empathy. In L. F. Barrett, M. Lewis, & J. M. Haviland-Jones (Eds.), *Handbook of Emotions*, Fourth Edition (pp. 871–885). Guilford Publications.

Revista Română de Filosofie Analitică,
Romanian Journal of Analytic Philosophy
Vol. XVII, Iulie-Decembrie 2023, Nr. 2, p. 32-61
ISSN (ediția electronică): 1843-9969,
ISSN (ediția tipărită): 1844-2218
DOI: 10.62229/rrfaxvii-2/2

AGENCY AFTER THE FACT: SELF-DECEPTION AS RETROSPECTIVE EXPLANATION

ELENA GIORGIANA ROŞU[1]

Abstract: The limitations of both intentionalist and motivationalist accounts suggest that the central difficulty in theorizing self-deception does not lie in whether intention is present, but in how beliefs are reorganized and reassessed within the agent's broader epistemic framework, and in how these beliefs are employed by explanatory reasoning. Intentionalist models over-intellectualize self-deception by positing goal-directed strategizing, while motivationalist accounts deflate the phenomenon by redescribing it as biased belief formation assessed against an external standard of rationality. Both approaches are naturally read as presupposing that agents operate on beliefs, in the first instance, with a primary aim towards truth, and that deviation from truth therefore requires special explanation, whether in terms of unconscious intention or motivational interference.

I propose instead that self-deception be understood against a more general account of belief reassessment, one that takes the agent's epistemic perspective as primary and treats explanatory coherence, rather than truth-tracking, as the organizing principle of belief organization. On this view, what requires explanation is not why agents sometimes fail to revise false beliefs in light of evidence, but how belief systems reorganize themselves to preserve stability while accommodating that evidence. This reframing allows us to retain the motivational insights of deflationary

[1] Elena Giorgiana Roşu is a master student in the 'Mind the Brain' programme, Faculty of Philosophy at the University of Bucharest.

accounts while explaining why patterns of belief reassessment are often interpreted, by agents and observers alike, as intentional or strategic only in retrospect.

Keywords: agency, motivated reasoning, valued beliefs, self-deception, retrospective explanation.

Introduction

The phenomenon of self-deception occupies an uncomfortable position in the philosophy of mind. It is familiar enough to be taken as a datum[2], yet it resists characterization as either a standard epistemic failure[3] or a deliberate act of self-misleading.

In this text, I propose that the difficulty might arise from a shared assumption underlying existing accounts: that agents are primarily oriented toward truth, and that deviation from truth, therefore, requires special explanation. Instead, I argue that self-deception is not a prospective epistemic failure at all, but a retrospective attribution an agent adopts once a valued belief has been reassessed within a coherence-preserving framework.

Existing accounts have approached the phenomenon from two directions. Intentionalist models treat self-deception as goal-directed: the agent or some partitioned subsystems deploy strategies to arrive at a desired belief while retaining some awareness of the truth. Motivationalist models deflate this by attributing biased belief formation to motivational factors operating below the level of intention. Both camps, despite their differences, assess the phenomenon from a third-personal standpoint. Neither explains the characteristic asymmetry of belief revision in terms available from the agent's own epistemic standpoint.

[2] We attribute it to others and recognize it retrospectively in ourselves.
[3] When one self-deceives, one is not simply mistaken.

An adequate account must explain the asymmetry first-personally. What makes certain beliefs resistant to revision while others remain negotiable is not their truth-value, but their functional role within the broader framework through which the agent organizes experience and generates explanations.

The text proceeds as follows. Section 1 develops a framework for everyday explanatory reasoning, arguing that individual belief systems, rather than truth, serve as the primary organizing structure for explanation. Section 2 discusses intentionalist and motivationalist accounts, identifying in each a shared reliance on external standards of rationality. Section 3 details an account of belief value and explanatory stability: beliefs differ in value as a function of their explanatory power, and this asymmetry predicts which beliefs are protected and which are open for reassessment under evidential pressure. Section 4 examines how motivational factors shape explanatory reasoning from the bottom-up, as constitutive constraints rather than distortions. Section 5 weaves these threads together: the intention to self-deceive is self-ascribed after the fact, as agents make sense of their own motivationally-constrained reasoning.

1. Explanation, Understanding, and the Role of Belief Systems

Hempel and Oppenheim (1948) have famously characterized explanations as bipartite, consisting of:

(1) an explanandum - the target phenomenon to be explained, and

(2) an explanans - the class of those propositions which are adduced to account for the phenomenon.

An explanation, under their deductive-nomological (DN) model, would be a valid deductive argument wherein the explanandum follows as a conclusion of the premises in the explanans. Therefore, the explanation should describe a logical relation between premises and

conclusion, in which the former shows why the latter obtained (Salmon, 1990, 2006, p. 7).

The authors note that an explanation is adequate insofar "its explanans, if taken account of in time, could have served as a basis for predicting the phenomenon under consideration," and thus conclude that the role of a scientific explanation is "not merely to record the phenomena of our experience, but to learn from them, by basing upon them theoretical generalizations which enable us to anticipate new occurrences and to control, at least to some extent, the changes in our environment." (Hempel & Oppenheim, 1948, p. 138).

To fix terms, in what follows, explanations necessarily operate on and with beliefs.[4] It could be said that one's belief system is the totality of classes of propositions that can be adduced to account for phenomena. An explanandum triggers explanatory demand. For the demand to be satisfied, an explanans, or a class of beliefs, needs to be arranged in a deductively valid structure from which the explanandum obtains. This is how one resolves the explanatory demand.

It has already been argued that the function of the explanation is not to support the truth of its conclusion or premises – their truth is already presupposed when the explanation is accepted (Hempel & Oppenheim, 1948, as discussed in Salmon, 1990). The truth of a belief, understood as a property of its corresponding proposition relative to the world, is assumed in cases where accepted explanations recruit this belief. What's left for the explanation under this model is to facilitate understanding of both the target phenomenon and the relations that obtain in the class of beliefs held as a broader, unified explanans. Explanation and understanding have long been characterized as two sides of the same coin, or at least as strongly implying each other (Grimm, 2010; Hills, 2016; De Haro & Butterfield, 2025).

[4] Importantly, I aim to stay neutral about the ultimate metaphysics of what beliefs are (for discussion, cf. Dumitru, 2004). In this text, I only approach the roles beliefs might play in self-deception.

More recent attempts to characterize explanations appeal to causes as the reasonable explanations for events (Salmon, 1984; Strevens, 2008; Woodward, 2003).

In parallel, research in psychology (Einhorn & Hogarth, 1986) has put forward evidence that explanations, in both science and everyday inference, typically appeal to causes. Furthermore, knowledge of general causal patterns limits which causes are judged probable (Einhorn & Hogarth, 1986) and relevant (Lombrozo & Carey, 2006; Lombrozo, 2006) during explanatory reasoning. For our purposes, it suffices to assume that at least a subset of explanations follows a causal structure, and that explaining via causal relations is ubiquitous in everyday explanatory reasoning (I did x because y; q happened because r).

For example, in explaining why the lawn is wet, one picks out a cause (i.e., rain), presupposes a general causal pattern under which the explanandum falls (i.e., rain causes surfaces on which it falls to become wet), and determines which parts of the phenomenon's causal history are relevant for explanatory purposes (but…one was on the lawn a minute ago and it was dry!)

Scientific explanations prioritize truth as a norm. However, in everyday explanatory reasoning, that normative framework no longer exclusively guides explanatory efforts. I contend that everyday explanations are functions of the broader belief systems individuals possess. An agent's belief system plays as much of a role in structuring and organizing the explanation as the agent itself. The totality of beliefs and their relations accepted to date determines the entire explanatory arena within which one makes sense of oneself and the external world.

To illustrate this dynamic, I propose that the phenomenon of self-deception constitutes the ideal stress test. In both the classic issues proposed by early literature on self-deception and the current gaps left by the now-orthodox motivationalist account, certain features of explanatory reasoning help bring forward an agent-centered account that exposes the tension between truth, motivation, and internal stability.

2. Self-Deception and the Limits of Truth-Centered Accounts

2.1. Intentionalist Models and the Strategic Puzzle

Self-deception was traditionally modeled on interpersonal deception, where A intentionally gets B to believe a proposition p, while they themselves believe that ~p, with the intention that B acquires or maintains the false belief p. Self-deception, modeled as such, raises the classic problem – already explicit in Davidson (1985, 2004) – that intentional self-deception seems paradoxical.

Two different paradoxes have been raised against the traditional model. The agent must be in a seemingly impossible state of mind wherein they genuinely hold, in full awareness, contradictory beliefs – a puzzle referred to as static or doxastic (see Mele, 2001, pp. 7–8). Moreover, it is necessary for the agent to intentionally deceive themselves without rendering their efforts futile – known as the dynamic or strategic puzzle (pp. 7–8).

Early literature was divided roughly into two main camps based on how they addressed the dynamic puzzle. The intentionalist camp maintained that self-deception is intentional and posited certain partitions that could assume the deceiver and deceived roles. Some intentionalists have argued, for instance, that self-deception is a temporally extended process, during which an agent having the true belief p consciously sets out to deceive themselves that ~p, then eventually forgets having had both the initial true belief and the intention to self-deceive (Sorensen, 1985; Bermúdez, 2000). The most prominent intentionalist accounts, however, proposed a mental partitioning of the self, where a subsystem or quasi-agent with varied degrees of rational agency and responsibility assumes the deceiver role.

Thus, intentionalist accounts strive to preserve a sense of agency, strategy, or control the agent has over self-deception episodes, but they overcommit on partitioning, unconscious homunculi, and certain goal

posits. To illustrate, a recent attempt to revive the intentionalist position belongs to Funkhouser and Barrett (2016), who propose that the unconscious plays the role of the deceiver. This conscious-unconscious split has been previously discussed by von Hippel and Trivers (2011), who maintain that the split better equips self-deceivers to deceive others.

Funkhouser and Barrett propose there are cases of "robust" self-deception where the unconscious can deploy certain strategies to accomplish a conscious – but not necessarily explicit – goal. Therefore, it is required that the agent has a goal to mislead themselves regarding some state of affairs. The agent must then jointly engage in:

"(1) the strategic pursuit of that goal [the goal to mislead themselves],

(2) in a way that is flexible to the nuances of possibly changing situations, and

(3) which involves some retention of the truth (or at least a non-trivial doubt)." (Funkhouser & Barrett, 2016, p. 683).

Regarding (1), the authors maintain: "Strategy requires a goal and rationality—strategizing is the rational pursuit of a goal. But we need not think of such rationality as conscious, deliberate, or ideal" (Funkhouser & Barrett, 2016, p. 683). Regarding (2), the unconscious is said to direct context-sensitive differentiated responses to shifting circumstances or evidence to avoid the truth better or perpetuate the falsehood. This condition already implies the last one – some awareness of the truth is preserved so that it can anchor deceiving behavior.

By invoking the agent's goal to mislead themselves regarding some state of affairs as the enabling condition for the unconscious to engage in robust self-deception, the authors are claiming some space between their account and the revisionist camp that originally deflated intention to motivated bias, and that consequently became the dominant model of self-deception.

If we renounce this goal condition and solely posit that there are such things as strategies, understood as cognitive patterns resulting from an individual's causal history that directly constrain explanatory reasoning, we can leverage their notion of unconscious to argue that, sometimes, sub-personal processes are organized according to certain patterns we might intuitively understand as deliberate. It is not that these patterns, in themselves, self-organize to pursue a goal, but rather that they have an inherent organization which we can explain by appealing to reasons. In ascribing them a goal, we grasp their underlying structure.

Importantly, the distinction between unconscious strategy and strategy-as-a-pattern that appears goal-directed lies in when the intention is ascribed. An unconscious strategy may be said to lead to a determined goal regardless of whether the goal is ever made explicit. In contrast, when an explicit goal explains a pattern as strategic, the goal is retrospectively attributed.

2.2. Motivational Bias and the Deflation of Intention

A separate camp has argued that we should remain skeptical of partitioning models and, instead, approach the phenomenon without appealing to "psychological exotica" (Mele, 2001). Non-intentionalist and deflationary approaches argued that most garden-variety cases of self-deception could simply be interpreted as "being mistaken" or "believing falsely." Yet simply acquiring a false belief, only to later reassess it when faced with contrary evidence, is a case of being mistaken, but it would not be classified as an episode of self-deception. In trying to account for this difference without appealing to intention, Mele (2001) has construed the self-deception phenomenon as a general category of motivationally biased judgment, effectively collapsing the revision of belief camp into a motivationalist account (von Hippel & Trivers, 2011; Lynch, 2012, 2013; Nelkin, 2002; Scott-Kakures, 2002; Levy, 2004).

In contrast to the intentionalist accounts, motivationalist approaches appear more intuitively attractive. On this view, biased belief formation or maintenance is motivated (by desire, fear, anxiety, or self-esteem protection), biases operate subpersonally, and there is no intention to deceive oneself. Agents do not aim to believe falsely; rather, self-deception minimally involves a person who "(a) as a consequence of some motivation or emotion, seems to acquire and maintain some false belief despite evidence to the contrary and (b) who may display behavior suggesting some awareness of the truth" (Mele, 2001; Deweese-Boyd, 2023). This definition seems to capture most garden-variety instances of self-deception without appeal to notions that might stir controversy. The motivationalist account is at least more parsimonious, which already provides a reason to more readily accept it.

It has been argued, however, that committing to this view prevents us from correctly picking out episodes of self-deception from other forms of motivated believing, such as wishful thinking (Bach, 1981) or self-delusions (Funkhouser & Barrett, 2016), as well as from explaining why motivation seems only selectively to produce bias (Bermúdez, 1997, 2000), or from capturing the characteristic 'tension' or internal conflict that accompany self-deception episodes (Nelkin, 2002).

2.3. The Epistemic Perspective Problem

In the context of my broader argument, I highlight a separate objection to a motivational account of self-deception. I contend that the account fails to capture cases in which the occurrence of self-deception depends on the agent's own epistemic perspective rather than on external assessment. By grounding self-deception in apparent resistance to "contrary evidence" and displayed behavior, the definition relies on an observer-relative standard of rationality, thereby treating self-deception as a second-person attribution rather than a genuinely first-person mental state. While

observers can *ascribe* self-deception, agents may never experience or recognize themselves as self-deceived. The classic objections raised against this account (e.g., tension) presuppose assessments made from the first-person perspective, yet the minimal definition allows instances that presuppose a second-person framework, leaving room for confusion.

When evidence threatens an agent's valued beliefs, reassessment is characteristically asymmetrical: some beliefs absorb the evidential pressure while others remain intact, even when the evidence equally bears on both[5]. Neither account explains this selectivity in terms available from the agent's own standpoint: the intentionalist attributes it to unconscious strategy, the motivationalist to motivational distortion, both of which are observer-relative descriptions.

An adequate account, therefore, must account for this asymmetry from the agent's own epistemic perspective, without presupposing that the agent's primary orientation is towards truth. It is not a belief's truth value that makes it negotiable or resistant to reassessment, but its functional role within the broader framework through which the agent organizes experience and generates explanations[6].

3. Belief Value, Meaning-Making, and Explanatory Stability

3.1. Meaning as Coherence in Explanatory Frameworks

To support the shift toward belief reassessment as stability-preserving[7], rather than truth-tracking, we must establish according to what references

[5] E.g., an agent who receives negative feedback on a project might readily reassess beliefs about the difficulty of the task, the competence of the evaluator, or the adequacy of the time available, while leaving intact a belief about their own ability, even when the evidence bears equally on all of them.

[6] See Thagard, (1989).

[7] It is not my goal to argue for stability as the goal of belief reassessment – more of an instrument among several. Stability can be subordinated to a deeper value it serves (e.g., good science). The same framework that predicts conservatism under normal evidential

this stability holds, therefore what value serves as a ranking factor in determining which beliefs are systematically protected, while others remain readily negotiable. It is often argued that one of the important instrumental functions of beliefs is to serve as meaning-making devices, thus allowing and supporting the construction of explanations for phenomena. Gopnik & Wellman (1994) and later theory-theory accounts of mindreading have framed beliefs as components of "naive" explanatory frameworks. Heider (1958) argued that people are fundamentally driven to explain actions and outcomes in ways that preserve coherence and predictability, therefore beliefs that successfully explain many events should be motivationally privileged. Beliefs, thus construed, should not be evaluated solely based on their truth value, but also relying on the extent to which they organize experience, sustain coherence, and render action intelligible.

The Meaning Maintenance Model (Heine et al., 2006) posits meaning as "what connects things to other things in expected ways." Meaning, relation, or association can be used interchangeably to refer to the output of an "innate capacity" people have, which they employ to "identify and construct mental representations of expected relationships between people, places, objects, and ideas." This capacity operates in three different domains: people seek coherent relations within the external world, within themselves, and within themselves and the external world. When an individual's sense of meaning is threatened in a domain (i.e., when they detect "structural breakdowns and inconsistencies," or are "otherwise confronted with meaninglessness"), they might engage in what the authors term "fluid compensation" to reaffirm meaning in frameworks alternative to that in which the threat occurred (pp. 89-91).

I propose explanations to be a special class of meaning relations. While not exhaustive of everything that can obtain as meaning, for an individual, accepting an explanation for x suffices for understanding x well enough (by that individual's standards). Some phenomena might

pressure also predicts and licenses radical revision when intermediate adjustment becomes too costly. I am grateful to Andrei Mărășoiu for pointing out how this could be perceived as an overcorrection in my framing.

trigger a more intense explanatory demand – some things require more urgent explanatory intervention. The individual, in turn, could be set to understand certain phenomena more in depth than others.

Some phenomena might only need superficial explanations for a while, until some relevant features, or related beliefs, are reorganized and consequently trigger a demand for deeper understanding. At all times, an agent maintains a kind of equilibrium[8] between what explanations it has accepted so far, which constitute candidates for reassessment, and what evidence it is yet to integrate. This equilibrium can be conceived as an internal coherence (Thagard, 2000). Laurence BonJour, a leading defender of coherentism, notes that a belief is justified not by being related to something outside the system of beliefs, but by its coherence with the rest of that system (1985).

The truth of each belief, understood as a property it has in relation to the world, is less important for preserving this equilibrium than its relations with related beliefs. Drawing from these observations, I argue that meaning-making is a function aimed at maintaining coherence, explanations are a special class of meaning-relations, while beliefs are constitutive elements of explanations. Therefore, the relations that hold between beliefs so that they can serve explanatory purposes, and thus contribute to meaning-making, provide these beliefs with a value other than truth, which determines their status in one's broader explanatory framework.

3.2. Valued Beliefs as Explanatory Anchors

This property of beliefs has been proposed by Preston and Epley (2005), who examined "valued beliefs" as high-level commitments that serve a primarily explanatory role within the broader belief system by anchoring

[8] This equilibrium can be conceived as an internal coherence (Thagard, 2000). Laurence BonJour, a leading defender of coherentism, notes that a belief is justified not by being related to something outside the system of beliefs, but by its coherence with the rest of that system (1985). I will broadly rely on these authors for the notion of coherence used in this text.

multiple explanations and organizing diverse observations[9]. The perceived value of a belief depends on its explanatory power. When a belief is applied to explaining many observations, its value increases, while explaining the belief itself (i.e., reducing it to underlying causes) decreases its value.

This stance has been confirmed in three experiments. Participants were introduced to a novel scientific finding, asked to evaluate other people's beliefs, or their own religious beliefs. In all scenarios, participants were assigned to either list observations the belief could explain (application condition) or list reasons why the belief might be true (explanation conditions).

Across experiments, the authors found a strong effect of the experimental condition on the belief's perceived value, but no effect on the belief's perceived truth. The effect was strongest when participants listed many applications, which shows that value increases with explanatory breadth. The same strong effect was found when participants operated with other people's beliefs, indicating that the mechanism is not limited to self-relevant beliefs. Moreover, beliefs were especially resilient and valuable when they were easy to apply but hard to explain. Religious beliefs, for example, resist reductive explanations, but are highly applicable as explanations, and so they are often fiercely defended.

A paradigmatic case of a highly resilient belief is detailed by Wegner (2002), who discusses free will as illusory. He argues that people have an "ideal of conscious agency" (p. 173) that guides their inferences and allows them to self-ascribe intentions over their actions, even when those actions could not have been intended. Many unintended behaviors we perform, he notes, require "some artful interpretation to fit them into our view of ourselves as conscious agents" (idem). This effectively describes a highly valued, especially resilient belief, which anchors explanations for one's and others' actions. The "artful interpretation" could be conceived as a pseudo-epistemic[10] reorganization of related

[9] I am grateful to Sandra Brânzaru for pointing out this line of research.

[10] That reorganization is pseudo-epistemic because the agent appears to themselves as sincerely pursuing the truth of the explanation.

beliefs through which one integrates disconfirming evidence without adjusting the valued beliefs.

For example, suppose Dan is committed to the truth of the proposition: "I consciously will my actions." When Dan encounters neuroscientific evidence showing that neural activity predictive of action (e.g., readiness potentials) reliably occurs before he reports a conscious decision to act (classically associated with experiments following Benjamin Libet), he does not readily discount his belief. He will, instead, reinterpret some constitutive beliefs, such as one stating that conscious action must be the earliest causal event in action initiation. He could now hold that action might be initiated unconsciously, but that he consciously endorses, controls, inhibits, or vetoes those actions. Therefore, the core belief is protected – he can still believe he consciously wills his actions, just not in a temporally primitive way.

3.3. Intermediate Beliefs and Predictive Stability

The process through which the broader belief system can be reorganized to accommodate evidence while preserving the perceived truth of a valued belief is exemplified in experiments by Wentura and Greve (2003; 2005). The authors argue that people protect (or immunize) their self-concept by preferentially reassessing some intermediate beliefs. When forced to acknowledge evidence that threatens personal, desirable traits, people integrate the evidence by reassessing more negotiable supporting assumptions. For example, consider the following set of beliefs:

(1) "I am erudite."

(2) "Good knowledge of history is necessary for being an erudite person."

(3) "I have good knowledge of history"

These refer to an individual's belief that they possess a general trait they find desirable (1), the belief that a particular skill is highly diagnostic for the trait (2), and the belief that the individual possesses the skill (3).

Students who held these beliefs took a difficult history test alongside an accomplice who knew the answers in advance. The accomplice predictably received an excellent score while the students failed. The evidence directly threatened (3), and indirectly threatened (1) via (2). But rather than adjust (1), which was rated as a desirable trait, participants preserved it by adjusting the diagnostic value of the skill, thus giving lower ratings to (2). This "peripheral adjustment" (Greve, 2010, p. 722) maintains the stability of the self-concept when faced with developmental changes and losses without completely disregarding reality.

Although Wentura and Greve focus on self-relevant beliefs, similar buffering mechanisms appear to operate more broadly across belief systems. Marchi and Newen (2022) have framed this phenomenon in predictive processing terms and purportedly expanded it for beliefs unrelated to one's self-concept. They argue that many of our beliefs are not immediately defeasible by perceptual evidence, but are instead connected to observable evidence via a set of intermediate beliefs. These intermediate beliefs might specify diagnostic criteria, causal pathways, or situational assumptions that connect abstract, high-level commitments to concrete evidence. They are flexible and allow adjustments to preserve internal consistency without discounting evidence. In this sense, it could be said that intermediate beliefs absorb prediction error so that valued beliefs remain stable (Marchi & Newen, 2022; Friston, 2010), unless the pressure becomes overwhelming or systematic.

If we describe this pattern more formally from a predictive processing perspective, valued beliefs resemble high-level priors with broad explanatory scope. These priors are slow to update and resistant to local prediction error, as they are justified by long-term coherence rather than immediate evidence (Friston, 2010; Clark, 2013). Under this framing, we can apply the above findings, which relate primarily to self-relevant beliefs, to one’s broader belief system.

Taken together, these findings converge on a shared architecture for belief systems. Sense-making implies a coherence-based process (Thagard, 2000) of belief reassessment, where explanations are functional outcomes. Such reassessment, however, is only possible against a relatively stable frame of reference. If all beliefs were equally vulnerable to reassessment at all times, no belief could function as an explanatory standard. Beliefs become increasingly valued as their breadth of application grows and as they resist reductive explanation, allowing them to function as explanatory anchors rather than as candidates for ongoing revision. The sustained commitment to such valued beliefs provides a fixed frame against which evidence can be interpreted and redistributed.

Intermediate beliefs, in contrast, are expendable commitments that interface between valued beliefs and the empirical world. Their flexibility allows the belief system to remain dynamically stable. When evidence threatens valued beliefs, its implications are selectively redistributed and absorbed by intermediate beliefs. Thus, the disconfirming evidence is accommodated, rather than wholesale ignored, denied, or repressed, but valued beliefs remain intact, and coherence is preserved.

3.4. Commitment Costs and Asymmetric Belief Revision

One important axis of belief value is normative and practical, rather than just explanatory. It is not solely the belief's explanatory value that constrains resistance to reassessment or reduction, but also a set of practical, inferential, and evaluative commitments that follow from representing a belief as true. Going back to the experiment proposed by Wentura and Greve (2003, 2005), accepting that "good knowledge of history is necessary for being considered an erudite person" means, *inter alia*, committing to denying erudition to those who lack historical knowledge. Reassessing this belief leads to changes in how one judges oneself and others (i.e., who counts as erudite, what counts as intellectual failure), as well as changes in how one is disposed to act, respond or reason going forward.

In contrast, accepting "I am not erudite" implies more costly commitments, depending on the trait's centrality to one's self-concept and broader circumstances. For example, one might be forced to find different explanatory premises for their past academic success, and recalibrate downward predictions about their future academic performance. They must accept a lower intellectual status among others they've previously considered peers, accept feelings of inadequacy, disappointment or diminished self-esteem as appropriate, and perhaps owe intellectual deference to people they previously considered inferior.

These commitments are rarely made explicit, but might be revealed when valued beliefs are put under evidential pressure. To accept a new belief implies accepting its commitment profile. The more a belief is valued, the more we can expect dismissing it to imply a more costly commitment profile, as it requires reworking a wide range of dependent explanations and evaluations.

The scope and cost of these commitments help further explain why agents preferentially reassess intermediate beliefs when available. The asymmetry is as much epistemic (some beliefs are harder to revise because of their explanatory power) as it is motivational (some beliefs are harder to revise because they incur higher commitment costs). Since these commitments can become explicit under pressure, it could be argued that they provide some access to what the agent might construe as reasons for why some of their beliefs are privileged.

Consider the following example[11]. When a mother whose son is missing asserts “My son is not dead,” she does so in response to subversive or overt external pressure to accept the contrary. The police officer might have told her that the likelihood of him being found alive decreases each day. She might have heard of an unfortunate example in her extended social network. She was likely exposed to several such stories in the media. However, after several years, she refuses to accept that her son is dead and braves any kind of persuasion attempt by saying, “I’ll only believe it if I see it with my own eyes!”

[11] I am grateful to Daniel Hutto for providing this example, among many others that helped shape this account into its final form.

It is, *prima facie*, difficult to deny that some awareness of the high likelihood of her son's death is available to her. That possibility must have been entertained as an explanation for why he still has not returned home after all these years, for instance. But she entertains that hypothesis only until she declares, and thus reinforces, the only condition under which the hypothesis will be accepted: only when she sees it with her own eyes. It is not the case that she completely resists evidence according to which her son might be dead, but she explains the hypothesis away by reassessing intermediate beliefs. For instance, she might convince herself that it would actually be easy for her son to survive all these years because he was exceptionally resourceful for a child his age.

By establishing a certain condition that needs to be satisfied before she can accept the belief, she immunizes the belief according to which her son is alive. The commitments that follow from accepting the contrary belief are deferred. Importantly, from her own perspective, the episode could only be considered self-deceptive when she eventually sees her dead son with her own eyes and finally accepts the belief. If her son miraculously turns out to be alive, then it could reasonably be said that she had just been *hoping* all along[12]. There are certain interpretations of this scenario as wishful thinking or even self-delusion. I propose the following distinction:

It is only self-deception insofar as self-deception is a reasonable explanation for the phenomenon. If the son turns out to be alive, then the mother's "hoping" requires no explanation. Nor does wishful thinking. If the son turns out to be dead, self-deception is an explanation given for why the mother failed to reassess the false belief in light of the evidence. Given that the belief she was defending against was ultimately accepted, the evidence that was so far deferred to intermediate beliefs now counts against the previously held belief. This signals an error that must be explained itself. Self-deception is the preferred explanation for such cases

[12] Suppose later on she learns of his death and subsequently reconceives death as afterlife so as not to accept that her child simply is no more. As beliefs about the afterlife are commonly held as unfalsifiable, this isolated scenario would not, of itself, constitute an instance of self-deception in the sense I discuss. Thanks to Andrei Mărășoiu for the question.

because it follows a broader explanatory strategy we employ – we often retroactively ascribe intention to our own and other people's behavior, by excellence.

The explanation is so easily accepted because motivation determines the selection space in which hypotheses are activated, entertained and evaluated. If our explanatory reasoning could be said to follow certain strategies, then the goal of those strategies can be attributed *ad hoc*. It is the structural patterns these strategies form that one interprets as directional, as meant to, or intending, to satisfy a goal, accomplish a desire, or protect a valued belief. But such structural patterns[13] derive from the same organization of beliefs that allows agents to identify and reflect on their desires, fears, anxieties and broader overall motivational factors.

4. Motivated Explanation and the Structure of Explanatory Reasoning

4.1. Two Phases of Explanatory Reasoning

I've described how belief systems are organized according to some coherence-preserving and practical principles. The following section looks at how beliefs operate in real-time reasoning within the broader process of explanation generation and evaluation. Consider the following account of motivated explanation, provided by Patterson et al. (2015), who (non-exhaustively) list the cognitive processes involved in the two distinct, but potentially overlapping phases of explanatory reasoning.

The first phase concerns explanation-generation and involves (1) the activation of candidate hypotheses on what the authors describe as "intuitive judgement on criteria for what qualifies as explanatory" (p. 2). Episodic and semantic memory search (2) then prompts the retrieval of

[13] I take such structural patterns to be recurrent configurations in how beliefs are activated, connected and weighted relative to one another (e.g., the consistent preferential weighting of hypotheses that preserve a central self-concept).

events, patterns and prior explanations relevant to the target of explanation. In cognitive updating (3), one manipulates the information held in working memory, including searching for new pro or con considerations, reassigning weights, evaluating credibility thresholds, judging coherence with background knowledge, and even reinterpreting old memories.

The second phase concerns the evaluation of candidate hypotheses and recruits all or a subset of the following processes: weighing evidence, judging coherence with background knowledge, judging simplicity, credibility, breadth, and depth. The authors discuss how all processes involved in generation and evaluation are points of vulnerability, where biases, heuristics, and motivational influences can intervene, leading the agent to some preferred explanation, in disfavor of explanations that would more closely match epistemic standards – such as truth, or accuracy.

4.2. Pre-emptive Constraint and Hypothesis Selection

Suppose I have noticed a friend has not replied to my messages for a few days. The event violates an expectation derived from prior regularities (i.e., they usually reply within a few hours) and triggers explanatory demand. I selectively activate plausible hypotheses, with existing beliefs constraining which hypotheses are even considered, and how they're further evaluated.

For example, I might entertain whether they are busy or upset with me, but the possibility that they have not seen my message does not occur to me, given an inductively established pattern in which my friend checks their phone frequently. This inference is implicit: the hypothesis simply does not register as a salient candidate explanation. However, if challenged – for instance, if my mother suggested that my friend might not see the message – I could make explicit the inferences that render the hypothesis implausible. The inference, therefore, while not consciously deployed, is retrievable when called upon to justify an explanation.

These implicit inferences are ubiquitous in everyday explanatory reasoning. What becomes explicit during this process has already been filtered through a motivationally biased lens (in this instance, perhaps hindsight or a confirmation bias might be operative). The example comes to show that the entire process is preemptively constrained by one's existing beliefs – what one already knows.

4.3. Motivation Without Distortion

The claim by Patterson et al. (2015), according to which motivation distorts reasoning, mirrors an assumption typically employed by motivationalist accounts of self-deception. It is assumed that, under no influence from motivational factors, there would be no instance in which the agent avoids, conceals, detours, or otherwise experiences a metacognitive state in which the truth could be associated with the tension or conflict characterizing self-deception. In contrast, when motivational factors exist, judgments become biased and might lead to or maintain false beliefs. The agent, therefore, is said to accept a false belief to satisfy some motivation. Evidence to the contrary constitutes a threat, the (at least partial) awareness of which would cause tension.

This stance, again, presupposes an external perspective. From the agent's own perspective, there need be no distinction between reasoning and its motivated counterpart. The explanatory strategies one employs are directly influenced by one's existing beliefs. What constrains explanation-generation and evaluation are personal and subpersonal patterns – an agent's causal history and the habitual cognitive patterns that obtain from that history. Motivational factors are already embedded in what the agent knows. The established causal patterns, broader narratives, past explanatory failures and successes – these all constrain explanatory reasoning to some degree. It has previously been assumed these constraints work on a process that aims at truth, therefore the intervention of motivational factors would somehow deceive the agent and prevent them from reaching the truth. I argue these motivational

factors shape the entire process, which is aimed at internal coherence, where the perceived truth of the resulting explanations is only prioritized on a case-by-case basis.

What an external observer might understand as subpersonal processes influenced by personal motivating factors resulting in a distortion of the truth, the agent understands as their personal search for meaning. When individuals commit to explanations, they assume their truth by organizing relations in their broader belief system so as to accommodate that truth. Therefore, the agent never commits to an avowedly false belief; they reorganize frameworks of related beliefs so that the truth of the target belief makes sense. The explanatory strategies one employs are as much intuitively (or unconsciously, subpersonally) fixed by the agent's causal history as they outwardly (consciously, personally) present themselves as a sincere pursuit of truth. But the explanations one accepts as true mark (1) a sufficient degree of understanding, judged from the agent's standpoint, and (2) coherence with the broader explanatory framework. The explanatory power of certain beliefs determines their value, which is why truth, understood as a property of propositions relative to the world, typically coincides with the "truth" an agent takes as a marker of acceptance for certain beliefs.

5. Self-Deceptive as Retrospective Attribution

Taken together, the foregoing account of motivated explanation clarifies how belief reassessment operates from the agent's own epistemic standpoint. What appear, from an external perspective, as biases intervening in otherwise truth-directed reasoning are, for the agent, integral constraints that shape which hypotheses are generated, which explanations are taken seriously, and which revisions are practically viable. Motivational factors do not enter explanatory reasoning as distortions imposed upon a neutral process; rather, they are already embedded in the organization of the belief system that governs explanatory coherence.

This reframing dissolves the need to treat self-deception as a prospective failure of rational belief management. If explanatory reasoning is structured around maintaining coherence within a valued framework, then the selective reassessment of beliefs under evidential pressure is not experienced as deception, but as sense-making. The appearance of intentional misdirection arises only when such patterns are retrospectively interpreted, by the agent or by observers, against a norm that treats truth as the sole regulative standard. With this in place, we can now reassess what self-deception amounts to under a coherence-centered account of belief reassessment.

Drawing on all previous observations, I argue that there are no episodes in which an agent is prospectively self-deceiving. Motivationalist accounts posit that purported motivations (desires, fear, anxiety, self-esteem protection) cause the agent to preferentially accept a false belief, even or especially when disconfirming evidence is available. It might be said that motivation effectively hijacks reasoning.

On my view, beliefs are reassessed preferentially as a fundamental function of explanatory reasoning. Some beliefs serve as high-level explanatory anchors, while others operate as flexible intermediaries. Under evidential pressure, flexible beliefs are reassessed so that evidence is accommodated without affecting the high-level explanatory anchors. A belief's value within the broader explanatory framework determines whether it is readily reassessed or protected via pseudo-epistemic reorganization of intermediate beliefs. In this view, belief reassessment is not truth-seeking, but stability-maintaining. Truth isn't the end goal – internal coherence is. These two goals often overlap, as in many cases, what is true is valuable in an explanatory sense. But treating them as interchangeable creates an unnecessary obstacle for accounts of self-deception.

It has been argued that the various cognitive processes involved in explanation-generation and evaluation are vulnerable to biases (Patterson et al., 2015). I propose that motivational factors effectively shape the entire explanatory space, from hypothesis generation, via confirmation and availability bias, to self-serving, selective memory retrieval, to biased

evidence weighting, asymmetric skepticism, belief perseverance, and so on. Rather than motivational judgements leading people to accept false beliefs as true, they limit which beliefs can be considered at all, in order to maintain the stability of the broader system.

Importantly, the system is open to reorganization. Previously implicit inferences can become explicit when they are needed to justify some explanations that the agent aims at, constrained by different motivational factors. Sometimes, accepting a hypothesis as an adequate explanation for a phenomenon requires reassessing a previously held valued belief. Sometimes, the valued belief is no longer valued because the agent was able to reduce it to its underlying causes. This triggers a broader reorganization in which evidence that was previously redistributed to some intermediate beliefs can be reinterpreted and held against a previously valued belief, as the threat it poses to the system's coherence is no longer relevant. These instances can themselves be explained first-personally by appeal to the interpersonal dynamic of deception. In ascribing agency to the patterns identified in one's subpersonal processes, agents are enabled to protect their ideal of conscious agency. Retrospective attribution of self-deception then serves as some form of scaffolding to identify patterns and make sense of one's reasoning. The structure, which appears deliberate, emerged incidentally.

6. Conclusion

In the proposed belief reassessment account, self-deception is not a distinctive mental act that precedes belief revision nor a motivational intrusion that derails an otherwise truth-directed process. It is, instead, a retrospective explanatory stance adopted once a belief has been reassessed within a coherence-preserving framework. What motivationalist accounts treat as the cause of epistemic error emerges here as a consequence of how agents make sense of their own reasoning when valued beliefs are eventually relinquished or exposed. By shifting the burden from truth-violation to stability-maintenance, my view dissolves the need to assume an external-observer perspective to account for tension, as well as

the need to posit prospective self-deception by appealing to intentions and partitions, yet preserves the felt tension between what one comes to believe and how one understands oneself as a rational, self-governing agent.

Acknowledgements

An earlier version of this work was presented at the 'Theory or Narratives? New Grounds for the Theory of Mind' workshop, held October 11-12, 2025 in Bran; I am grateful to Sandra Brânzaru, Andrei Mărășoiu, Daniel Hutto, Zuzanna Rucinska and Daniel Stancu for ample discussion and feedback. Any remaining errors are my own.

References

Bach, K. (1981). An Analysis of Self-Deception. *Philosophy and Phenomenological Research*, *41*(3), 351. https://doi.org/10.2307/2107457

Baghramian, M., & Nicholson, A. (2013). The Puzzle of Self-Deception. *Philosophy Compass*, *8*(11), 1018–1029. https://doi.org/10.1111/phc3.12083

Balcetis, E., & Dunning, D. (2010). Wishful seeing: More desired objects are seen as closer. *Psychological Science*, *21*(1), 147–152. https://doi.org/10.1177/0956797609356283

Bargh, J. A., & Morsella, E. (2008). The Unconscious Mind. *Perspectives on Psychological Science*, *3*(1), 73–79. https://doi.org/10.1111/j.1745-6916.2008.00064.x

Barnes, A. (1998). *Seeing through Self-Deception* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511583353

Bermudez, J. L. (1997). Defending intentionalist accounts of self-deception. *Behavioral and Brain Sciences*, *20*(1), 107–108. https://doi.org/10.1017/S0140525X97270032

Bermudez, J. L. (2000). Self-deception, intentions, and contradictory beliefs. *Analysis*, *60*(268), 309–319. https://doi.org/10.1111/1467-8284.00247

BonJour, L. (1985). *The structure of empirical knowledge*. Harvard University Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Cooper, M. L., Agocha, V. B., & Sheldon, M. S. (2000). A Motivational Perspective on Risky Behaviors: The Role of Personality and Affect Regulatory Processes. *Journal of Personality*, *68*(6), 1059–1088. https://doi.org/10.1111/1467-6494.00126

Davidson, D. D. (2004). *Deception and division*. https://api.semanticscholar.org/CorpusID:151767930

De Haro, S., & Butterfield, J. (2025). Understanding and explanation. In S. De Haro & J. Butterfield, *The Philosophy and Physics of Duality* (1st ed., pp. 531–551). Oxford University Press, Oxford. https://doi.org/10.1093/oso/9780198846338.003.0015

Deweese-Boyd, I. (2023). Self-deception. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Fall 2023). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2023/entries/self-deception/

Doody, P. (2017). Is there evidence of robust, unconscious self-deception? A reply to Funkhouser and Barrett. *Philosophical Psychology*, *30*(5), 657–676. https://doi.org/10.1080/09515089.2017.1328491

Dumitru, M. (2004). Atitudini propoziţionale. Probleme şi teorii. In *Explorări logico-filozofice* (pp. 204–243). Humanitas.

Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin*, *99*(1), 3–19. https://doi.org/10.1037/0033-2909.99.1.3

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Funkhouser, E., & Barrett, D. (2016). Robust, unconscious self-deception: Strategic and flexible. *Philosophical Psychology*, *29*(5), 682–696. https://doi.org/10.1080/09515089.2015.1134769

Gopnik, A., & Wellman, H. M. (1994). The theory theory. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the Mind* (1st ed., pp. 257–293). Cambridge University Press. https://doi.org/10.1017/CBO9780511752902.011

Greve, W. (Ed.). (2005). *The adaptive self: Personal continuity and intentional self-development*. Hogrefe & Huber.

Greve, W., & Wentura, D. (2003). Immunizing the Self: Self-Concept Stabilization Through Reality-Adaptive Self-Definitions. *Personality and Social Psychology Bulletin*, *29*(1), 39–50. https://doi.org/10.1177/0146167202238370

Greve, W., & Wentura, D. (2010). True lies: Self-stabilization without self-deception. *Consciousness and Cognition*, *19*(3), 721–730. https://doi.org/10.1016/j.concog.2010.05.016

Grimm, S. R. (2010). The goal of explanation. *Studies in History and Philosophy of Science Part A*, *41*(4), 337–344. https://doi.org/10.1016/j.shpsa.2010.10.006

Harmon-Jones, E., Harmon-Jones, C., & Levy, N. (2015). An Action-Based Model of Cognitive-Dissonance Processes. *Current Directions in Psychological Science*, *24*(3), 184–189. https://doi.org/10.1177/0963721414566449

Heider, F. (1958). The naive analysis of action. In F. Heider, *The psychology of interpersonal relations*. (pp. 79–124). John Wiley & Sons, Inc. https://doi.org/10.1037/10628-004

Heine, S. J., Proulx, T., & Vohs, K. D. (2006). The Meaning Maintenance Model: On the Coherence of Social Motivations. *Personality and Social Psychology Review*, *10*(2), 88–110. https://doi.org/10.1207/s15327957pspr1002_1

Hempel, C. G., & Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of Science*, *15*(2), 135–175. https://doi.org/10.1086/286983

Hills, A. (2016). Understanding Why. *Noûs*, *50*(4), 661–688. https://doi.org/10.1111/nous.12092

Hirstein, W. (2005). *Brain Fiction: Self-deception and the Riddle of Confabulation*. MIT Press.

Huang, J. Y., & Bargh, J. A. (2014). The Selfish Goal: Autonomously operating motivational structures as the proximate cause of human judgment and behavior. *Behavioral and Brain Sciences*, *37*(2), 121–135. https://doi.org/10.1017/S0140525X13000290

Keil, F. C. (2006). Explanation and Understanding. *Annual Review of Psychology*, *57*(1), 227–254. https://doi.org/10.1146/annurev.psych.57.102904.190100

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Lazar, A. (1999). Deceiving oneself or self-deceived? On the formation of beliefs "under the influence." *Mind*, *108*(430), 265–290. https://doi.org/10.1093/mind/108.430.265

Levy, N. (2004). Self-Deception and Moral Responsibility. *Ratio*, *17*(3), 294–311. https://doi.org/10.1111/j.0034-0006.2004.00255.x

Loewenstein, G. (1996). Out of Control: Visceral Influences on Behavior. *Organizational Behavior and Human Decision Processes*, *65*(3), 272–292. https://doi.org/10.1006/obhd.1996.0028

Loewenstein, G. (2000). Emotions in Economic Theory and Economic Behavior. *American Economic Review*, *90*(2), 426–432. https://doi.org/10.1257/aer.90.2.426

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, *10*(10), 464–470. https://doi.org/10.1016/j.tics.2006.08.004

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, *99*(2), 167–204. https://doi.org/10.1016/j.cognition.2004.12.009

Lynch, K. (2012). On the "tension" inherent in self-deception. *Philosophical Psychology*, *25*(3), 433–450. https://doi.org/10.1080/09515089.2011.622364

Lynch, K. (2013). Self-Deception and Stubborn Belief. *Erkenntnis*, *78*(6), 1337–1345. https://doi.org/10.1007/s10670-012-9425-0

Marchi, F., & Newen, A. (2022). Self-deception in the predictive mind: Cognitive strategies and a challenge from motivation. *Philosophical Psychology*, *35*(7), 971–990. https://doi.org/10.1080/09515089.2021.2019693

Mele, A. R. (1992). *Irrationality: An Essay on Akrasia, Self-Deception, and Self-Control*. Oxford University Press.

Mele, A. R. (2001). *Self-Deception Unmasked*. Princeton University Press; JSTOR. http://www.jstor.org/stable/j.ctt7s4tg

Miller, D. T., & Ross, M. (1975). Self-serving biases in the attribution of causality: Fact or fiction? *Psychological Bulletin*, *82*(2), 213–225. https://doi.org/10.1037/h0076486

Nelkin, D. K. (2002). Self–Deception, Motivation, and the Desire to Believe. *Pacific Philosophical Quarterly*, *83*(4), 384–406. https://doi.org/10.1111/1468-0114.t01-1-00156

Patterson, R., Operskalski, J. T., & Barbey, A. K. (2015). Motivated explanation. *Frontiers in Human Neuroscience*, *9*. https://doi.org/10.3389/fnhum.2015.00559

Preston, J., & Epley, N. (2005). Explanations Versus Applications: The Explanatory Power of Valuable Beliefs. *Psychological Science*, *16*(10), 826–832. https://doi.org/10.1111/j.1467-9280.2005.01621.x

Salmon, W. C. (2006). *Four decades of scientific explanation*. University of Pittsburgh Press.

Scott-Kakures, D. (2002). At "Permanent Risk": Reasoning and Self-Knowledge in Self-Deception. *Philosophy and Phenomenological Research*, *65*(3), 576–603. https://doi.org/10.1111/j.1933-1592.2002.tb00222.x

Sorensen, R. A. (1985). Self-Deception and Scattered Events. *Mind*, *XCIV*(373), 64–69. https://doi.org/10.1093/mind/XCIV.373.64

Strevens, M. (2009). *Depth: An Account of Scientific Explanation*. Harvard University Press.

Talbott, W. J. (1995). Intentional Self-Deception in a Single Coherent Self. *Philosophy and Phenomenological Research*, *55*(1), 27. https://doi.org/10.2307/2108309

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, *12*(3), 435–467. https://doi.org/10.1017/S0140525X00057046

Thagard, P. (2000). *Coherence in Thought and Action*. The MIT Press. https://doi.org/10.7551/mitpress/1900.001.0001

Trivers, R. (2011). *The folly of fools: The logic of deceit and self-deception in human life*. Basic Books.

von Hippel, W., & Trivers, R. (2011). The evolution and psychology of self-deception. *The Behavioral and Brain Sciences*, *34*(1), 1–16; discussion 16-56. https://doi.org/10.1017/S0140525X10001354

Watkins, P., Vache, K., Verney, S., Muller, S., & Mathews, A. (1996). Unconscious Mood-Congruent Memory Bias in Depression. *Journal of Abnormal Psychology*, *105*, 34–41. https://doi.org/10.1037/0021-843X.105.1.34

Wegner, D. M. (2002). *The Illusion of Conscious Will*. The MIT Press. https://doi.org/10.7551/mitpress/3650.001.0001

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press. https://doi.org/10.1093/0195155270.001.0001

Woodward, J. (2005). *Making things happen: A theory of causal explanation*. Oxford University Press, Incorporated.

Woodward, J., & Ross, L. (2025). 20th century theories of scientific explanation. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Winter 2025). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2025/entries/scientific-explanation-20th/

Revista Română de Filosofie Analitică,
Romanian Journal of Analytic Philosophy
Vol. XVII, Iulie-Decembrie 2023, Nr. 2, p. 62-85
ISSN (ediția electronică): 1843-9969,
ISSN (ediția tipărită): 1844-2218
DOI: 10.62229/rrfaxvii-2/3

HOW DO WE GROUND 'GROUNDING'? WHY THE VECTOR GROUNDING PROBLEM REMAINS UNSOLVED

FLORIN COJOCARIU[1]

Abstract: On the view I develop, grounding is a phenomenological achievement of organisms — structured, concern-laden encounter with the world, observable in basic minds that connect to their environment *without* representational content. Language coordinates with this connection; it does not create nor implement it. I develop these views within the Pattern Recognition Unity (PRU) framework, distinguishing experiential Pattern-Constellations from Linguistic Pattern-Constellations and exploring how they coordinate. Contra Mollo and Milliere's (2025) view, I further argue that teleosemantics — a theory that earned its credibility on pre-verbal organisms selected against grounded functions like feeding and fleeing — loses its explanatory force when applied to systems selected against linguistic adequacy. The selection history of LLMs *borrows* its world-connection from human raters rather than *constituting* its own; its causal relation to *worldly features* is inherited ancestry, not the proximal coupling teleosemantics describes in the natural world. It follows that no evaluation conducted within the linguistic register can distinguish genuine grounding from sufficiently powerful linguistic mastery.

Keywords: *Grounding; LLMs; Consciousness; Explanatory Gap; Teleosemantics; Basic Minds.*

[1] Florin Cojocariu is a student in the 'Analytic Philosophy 'master programme within the Faculty of Philosophy at the University of Bucharest.

1. The Grounding Problem

1.1.The Standard Question

How does the *word*" cat" connect to actual cats? This question — the grounding problem in miniature — has generated three decades of debate in philosophy of mind, cognitive science, and now philosophy of AI. From Harnad's (1990) Symbol Grounding Problem and Searle's Chinese Room (Searle, 1980) to the current dispute over LLMs, the question tracks the same structure: on one side, a representation (word, symbol, internal state); on the other, the world (cats, fires, chairs); between them, a gap that the theory must bridge.

The problem persists because representations *qua* representations relate only to other representations. A word connects to other words. A definition connects to other definitions. An internal state described as "carrying information about cats" connects to *the description of cats*, not to *cats*. This is the representational merry-go-round: representations all the way down, that never reach the actual world they're representing.

The standard framing, then, treats grounding as *a property that representations need to acquire*. The word "cat" lacks something — a connection, a hook, a tether to the world — and the task is to identify what that something is and under what conditions it can be obtained.

1.2. The Best Available Disambiguation

(Mollo and Milliere, 2025) bring about progress by showing that "grounding" is tasked with too many jobs at once.[2] They distinguish five notions:

(i) Referential grounding: a representation connects to worldly entities or properties.

[2] For an anticipation of questions concerning how these different notions of grounding overlap, albeit in a different terminology, that of reference v. description for singular terms, cf. (Dumitru, 2004).

(ii) Sensorimotor grounding: a representation is connected to sensory or motor representations *inside the mind*.

(iii) Relational grounding: a representation is linked to other linguistic representations through inferential or distributional relations *inside the mind*.

(iv) Communicative grounding: representations are calibrated through linguistic practices between speakers to establish common ground.

(v) Epistemic grounding: a representation is linked to a knowledge base — a corpus of specific knowledge, formally or informally built (text or culture).

Of the five, only referential grounding direct hooks onto worldly entities. Mollo and Millière conclude, correctly, that this is the only notion relevant to the Vector Grounding Problem for LLMs. They argue that the other four either trade between representations alone, or presuppose the semantic content they would establish, or defer the question to a further structure that itself requires grounding.

The four non-referential notions are not failed *kinds* of grounding. They are concepts *about* grounding: linguistic descriptions, formulated in publicly articulable prose, of how representations might relate to *something*. Each names a relation between representations. Each can be stated, refined, debated. Each is a description rather than a direct hook to the thing described. Concepts about grounding are not grounding.

Referential grounding is supposed to be grounding itself, and that conditions: causal-informational relations, selection history, the function of tracking worldly features. Simply coming up with another concept about grounding — another publicly articulable description, connected inferentially to other descriptions in the teleosemantic literature—would only generate a fifth entry in Mollo and Millière's list.

The problem is structural: language can only produce concepts about grounding. It cannot produce grounding itself, because grounding, if real, is not a linguistic relation. Every attempt to specify what referential grounding is converts it into a description of what it is. The representational framework reinforces this limitation: representations are

items that bear content, content is evaluated for accuracy, accuracy is assessed in language. Everything the framework touches becomes a candidate for linguistic evaluation. It cannot describe the organism's pre-linguistic encounter with the world without converting it into a representation.

The debate cycles because it treats grounding as a property language must deliver, when grounding is something, organisms achieve and language coordinates with.

1.3. The Hidden Assumption

The standard framing contains a hidden assumption that makes the problem structurally unsolvable from within. The assumption is this: grounding is a property of representations — something language needs to acquire. The question "how does the word 'cat connect to actual cats?" places grounding in the linguistic domain. Its subject is the word. The word is what lacks grounding. The world sits inert, waiting to be hooked onto. The entire problematic runs from language outward: *symbol → ??? → world*.

Consider a cat tracking a mouse. The cat's mind connects to the mouse with precision: position, speed, trajectory are tracked; the body organizes around the encounter; salience and urgency radiate from the prey. No language is involved. No representations with truth conditions are in play. Yet the connection between mind and world is as tight as any the grounding debate discusses.

What should we call this? It is not grounding in the standard sense — there is no linguistic content to ground[3]. But it is a mind-world

[3] This aligns with the Radical Enactive Cognition (REC) programme of Hutto and Myin (2013), which argues that basic minds — minds without language — are world-directed but contentless. On their account, genuine semantic content with conditions of satisfaction emerges only when organisms participate in sociocultural linguistic practices. What REC does not provide is a positive structural account of what basic world-directedness consists

connection: a structured, stable, concern-laden engagement with a worldly entity, formed through the organism's history of encounter. Call it *basic grounding*: the pre-linguistic, pre-content connection between an organism and its world that the standard grounding debate presupposes.

Basic grounding is what makes linguistic grounding possible. The child who learns the word "cat" does not use the word to reach cats. She already reaches cats — through basic grounding built from world encounter-events. The word arrives and coordinates with a connection that is already in place. The child's accumulated cat-encounters — serial, *locus*-bound, laden with affect and concern — have formed a stable experiential pattern: an attractor basin integrating visual, tactile, auditory, motor, and affective dimensions of cat-encounter. This pattern is not a representation of cats. It is the sedimented structure of the organism's encounters with cats from its particular *locus* in the world. It exists before language. It exists in animals who will never have language. It is built through encounter, not through representation.

Grounding, properly understood, does not run from language to world. It runs from world to organism — basic grounding, the formation of experiential patterns through encounter — and then language coordinates with the result. The standard question inverts this order. By asking "how does the word connect to the world?" it implies that the problem begins with language and must be solved within language. But the organism is already on the world's side. The bridge language needs is not to the world but to the organism's experiential engagement with the world — an engagement that is subjective, pre-linguistic, and unavailable in the collective register of language.

---

in — what its units are and how it is organized. The experiential Pattern-Constellations ({X}) developed in Section 2 are offered as this structural unit: contentless but stable, phenomenologically real, empirically investigable. Content, on this account, emerges at $\mathcal{R}(\{X\},\langle x\rangle)$— in the coordination between experiential and linguistic patterns — which locates the emergence of content exactly where REC predicts: in the organism's entry into sociocultural-linguistic practice. PRU thus completes rather than competes with REC, providing the structural architecture for a claim REC makes but does not fully specify.

1.4. Grounding Redescribed

If grounding is not a property of representations but a phenomenological achievement of organisms, what does the grounding problem actually ask? Not: "How do words connect to the world?" But: "How does language coordinate with the experiential patterns organisms have already formed through worldly encounter?"

The first question is unanswerable within language because it locates grounding in language. The second question is tractable because it separates things that the first question conflates:

(a) The experiential pole: the organism's stable pattern of encounter, subjective, individual, formed pre-linguistically through serial, *locus*-bound, concern-laden events. This is what basic grounding produces. It is what the cat has with respect to the mouse and what the child has with respect to cats before she learns the word.

(b) The linguistic pole: the public, collective pattern of word-use, intersubjective, sustained through communal practice. This is what language *is*.

(c) Multifarious coordination: the learned couplings between (a) and (b) — the processes by which an individual's experiential patterns become coordinated with a community's word-use patterns. This is what reference *is*.

The grounding problem, in its standard form, collapses (a), (b), and (c) into a single question about representations. The result is that (a) — the subjective, phenomenological pole — becomes invisible. It is not denied; it is simply erased by a framing that has no place for it.

What follows? Evaluating whether LLMs are "grounded" requires a framework that can hold (a) and (b) apart — that can mark the experiential pole without converting it into a linguistic description, and mark the linguistic pole without letting it silently stand in for the whole. A meta-language is then needed: a notation equipped with devices that indicate the experiential pole without redescribing it, mark the linguistic pole as linguistic, and represent the coordination between them as a structured relation between different kinds of thing.

Mollo and Millière's argument is that LLMs can have internal states genuinely *about* the world — not just interpretable as such — provided those states reliably track worldly features and were shaped by a learning process (especially RLHF) that selected them for doing so. On their view, neither embodiment nor multimodality is required; the right kind of training history is enough. This is the strongest available case that LLMs achieve referential grounding and it operates entirely within (b), the linguistic pole. Their teleosemantic conditions describe (b)-type properties: distributional structure, functional roles, causal-chain descriptions. LLMs, which are (b)-type systems par excellence, naturally satisfy (b)-type conditions. The evaluation succeeds because evaluator and evaluated share the same register. The experiential pole (a) is not detected as missing because the framework has no way to detect it at all. This is not Mollo and Millière's fault. It is a structural consequence of asking a question about subjective experience using tools that can only describe collective, publicly articulable properties.

2. The PRU Notation as Meta-Language

The preceding section identified three things a meta-language must mark: (1) the experiential pole — subjective, individual, non-transmissible through text; (2) the linguistic pole — collective, public, transmissible through text; (3) coordination relations between them. I develop three corresponding notational devices within the Pattern Recognition Unity (PRU) framework[4].

{X} — Experiential Pattern-Constellation. The curly braces function as *notational quarantine*: they indicate the experiential pole without redescribing it in linguistic terms. The moment you write "the

[4] Cojocariu, Florin, Pattern-Recognition Unity (PRU): A Framework Specification A Meta-Language for Grounding, Reference, and the Experience–Language Interface (February 10, 2026). Preprint available at SSRN: https://ssrn.com/abstract=6285878 or http://dx.doi.org/10.2139/ssrn.6285878

experiential constellation of fire," you have produced a piece of language — a concept about {X}. {FIRE} resists this by pointing at the experiential pole the way an index finger points at a dog. It does not describe the dog. It does not deliver the dog. It marks what is experienced when encountering a dog, and the description stops here.

{X} is what the cat has with respect to the mouse — the stable, multi-modal, affect-laden attractor basin formed through the organism's history of encounter-events. It is what the previous section called basic grounding, now given a notational handle. {X} is always individual, always formed through encounter-events, always *locus*-bound. {FIRE} is *my* fire-constellation or *yours* — there is no view-from-nowhere {FIRE}.

⟨x⟩ — Linguistic Pattern-Constellation. The angle brackets mark the linguistic pole *as linguistic*. ⟨dog⟩ denotes the public pattern of how "dog" functions in language — its distributional profile, inferential relations, sentential regularities — and nothing more. The framing is broadly Wittgensteinian: ⟨dog⟩ is the meaning of "dog" insofar as meaning lives in the public practice of use.

⟨x⟩ is always collective, always public, always transmissible through text. ⟨dog⟩ is not mine or yours — it is the communal pattern of use. It is what LLMs learn.

$\mathcal{R}$({X}, ⟨x⟩) — The coordination relation. $\mathcal{R}$ marks that referential grounding is a *structured relation* between two different kinds of pattern. It is not a single hook from word to world. It is the learned coupling between an individual's experiential attractor basin and a collective pattern of word-use. Reference works because $\mathcal{R}$ holds — because my {DOG} and your {DOG}, despite being different, are each coordinated with the shared ⟨dog⟩.

$\mathcal{R}$ marks the coordination as a coordination — as a relation between two things the notation has formally separated. Without the separation, "reference" looks like a single, undifferentiated relation between word and world. With the separation, reference reveals itself as a bridge between something the notation can describe (⟨x⟩) and something it can only indicate ({X}).

Throughout this paper I'll adhere to the following simplifying convention: if context asks for it, by {X} I mean the totality of the pattern constellation an organism developed at a given time, something more precisely written as $\bigcup_i\{X_i\}$. By ⟨x⟩ we mean the totality of language pattern constellations a human possesses at the same moment, something more precisely written $\bigcup_i \langle x_i \rangle$.

The notation does not deliver {X} to the reader. No notation could. What it does is create *representational hygiene* — a discipline that prevents a specific conflation from occurring unnoticed: the conflation of the collective linguistic pole with the full structure of grounding.

At this stage, PRU is less than a formalism. A formalism derives results, generates theorems, performs calculations. PRU does not do that here. It is a meta-language: a language about the limitations of language.

But the notation is also more than a metaphor. The {X}/⟨x⟩ distinction has empirical traction and developmental predictions, both developed below. As I view it, it is not a suggestive analogy but a structural claim about the architecture of cognitive grounding.

For practical purposes, PRU also operates at the level of individual words, tracking how the same word-form functions in two modes:

• x^o: an object-word — a word-token functioning in coordination with a specific experiential constellation. "Dog^o" as uttered while encountering a particular dog. x^o labels {X} — it is the linguistic element integrated into the experiential pattern through encounter.

• x^c: a concept-word — a word-token functioning within linguistic practice. "Dog^c" as it appears in "Dogs are mammals." x^c operates within ⟨x⟩ — the autonomous space where words connect to words.

• $\mathcal{R}(x^o, x^c)$ Reference at the word level — the concept-word coordinating with the object-word, which anchors to {X}.

A child who has acquired only dog^o can point and say "dog!" but cannot answer "what is a dog?" except by ostension. A child with dog^c can answer: "an animal that barks." The concept-word operates in the autonomous linguistic space; the object-word is tethered to encounter.

As I see it, the sequence that builds this structure is:

$\{X\} \rightarrow \{X, x^o\} \rightarrow \{X, x^o, x^c\} \rightarrow \mathcal{R}(x^o, x^c)$

Stage 1 is basic grounding: the pre-linguistic experiential pattern. Stage 2 integrates a label through encounter-events. Stage 3 achieves dual function: the same word can operate in object-mode or concept-mode. Stage 4 is reference proper: the coordination between modes[5].

This sequence is irreversible. It must begin with experiential encounter and proceed through label-integration before achieving the dual-mode structure that makes reference possible. In contrast, LLMs begin and remain at Stage 3 — concept-word-mode mastery without the developmental history that builds it from encounter. They excel at $\mathcal{R}(x^c, y^c)$, coordinating concepts with concepts, because their training corpus is precisely a record of such coordinations. But no amount of concept-concept mastery produces the object-mode anchoring that grounds reference, because x^o requires a history of encounter that text cannot transmit.

3. Returning to Mollo and Millière

3.1. Reconstruction

We can return to Mollo and Millière's teleosemantic argument. Mollo and Millière draw on teleosemantic theories of representational content – (Shea, 2018), (Millikan, 2017), (Neander, 2017) - to specify two conditions for referential grounding:

1. Causal-informational[6]: Internal states must correlate with worldly features through causal chains.

[5] The first two stages are separated here for clarity but available evidence for how infants acquire language suggests that these are tightly intertwined.

[6] The phrase "causal-informational relations" appears twenty-five times in Mollo and Millière's original paper. I adopt their terminology throughout, but note that in several instances what is described as "causation" is better characterized as correlation — as when the authors illustrate the concept: "smoke is *correlated* with fire because it is *caused* by it."

2. Historical/selectional: Those states must have been selected to carry that information — must have the function of tracking worldly features. They argue LLMs satisfy these conditions through three pathways:

Post-training: RLHF (Reinforcement Learning from Human Feedback) introduces extra-linguistic norms (factuality, helpfulness). Human raters reward world-tracking outputs, establishing a selection history with extra-linguistic success conditions.

Pre-training: Next-token prediction implicitly selects for internal states that track worldly regularities. Mechanistic interpretability evidence (Othello-GPT, probing studies) shows LLMs develop representations that model extra-linguistic structure.

In-context learning: Mollo and Millière further argue that mesa-optimization during inference can establish transient world-involving functions without parameter updates. Since this case is even more clearly intra-linguistic than the other two, the critique that follows applies *a fortiori*.

Their conclusion: LLMs achieve referential grounding. Multimodality and embodiment are neither necessary nor sufficient. The argument is sophisticated and represents, in my view, the strongest available case for LLM grounding, however, it does not succeed. The reasons converge on a single structural point: the argument is conducted entirely within ⟨x⟩ — the collective, publicly articulable register that is precisely what the grounding problem says is insufficient.

3.2. Teleosemantics Embeds but Does Not Constitute Grounding

3.2.1. Where the theory earns its force

A frog sits at the edge of a pond. A fly crosses its visual field; retinal ganglion cells fire; the tongue strikes. The frog eats. A frog whose detector

---

The slide is not accidental; it reflects a deeper difficulty with importing teleosemantic vocabulary into LLM analysis, which I address in section 3.4.

misfires — at shadows, at BB pellets, at nothing — goes hungry. Over evolutionary time, only frogs whose detectors reliably tracked flies in their environment survived and reproduced.

This is the scenario that gives teleosemantics its explanatory force. The theory says: the frog's retinal state has the function of indicating prey because ancestors whose states performed that function were differentially selected. This yields normativity without mystery — the state is *supposed to* track flies, and when it fires at a BB pellet it *misrepresents*, because its proper function is fly-detection regardless of what caused the particular firing (Neander, 1991). Normativity and misrepresentation are naturalized in one move[7].

Tree rings carry information about age, yet nobody thinks tree rings have content. What distinguishes the frog is that the information its states carry is useful: useful for catching prey, for surviving, for reproducing. The world, so to speak, *trains* the organism, and it trains against the organism's essential functions. Selection that tracked irrelevant information, however faithfully, would be eliminated. The "teleo-" in teleosemantics does not apply to all purposes, but only to purposes constituted by the organism's embodied predicament, where getting things wrong risks death.

[7] One might object that the frog's fly-catching is itself a teleosemantic case: the tongue-strike has the function of catching flies, selected through evolution. We grant this — indeed, the main text uses this description. But the description operates from the theorist's ⟨x⟩ register: it attributes content, conditions of satisfaction, and normativity to the frog's states as a third-person redescription. Whether the frog's cognitive economy itself requires content-bearing states is precisely what is at issue between teleosemantic and radical enactivist accounts. Our point is that what makes the frog grounded — its dynamic embodied coupling to the fly through sensory transduction, motor readiness, and evolutionary history — is prior to and independent of the teleosemantic redescription. The redescription is accurate as far as it goes, but it presupposes the coupling rather than explaining it. See (Hutto and Myin, 2013, chs. 3–4) for a sustained argument that basic cognition is contentless, and (Thompson, 2010) for the life-mind continuity that makes basic grounding a feature of all organisms, not only linguistically sophisticated ones.

3.2.2. Where it was forged

Teleosemantics earned its credibility entirely on pre-verbal systems: frogs, bees, magnetotactic bacteria (Dretske, 1988), (Millikan, 1989)— organisms whose states are said to acquire content through direct environmental coupling, prior to and independent of language. The theory works for these cases precisely because selection under environmental pressure operates in a register independent of the register in which we *describe* environmenal coupling. That independence is what gives the theory its naturalistic force: (what teleosemantics identifies as) content is explained without presupposing language, meaning, or interpretation.

Human language was forged in the same crucible. Language evolved and develops in the context of coordinated action: shared warning, tool use, teaching, negotiation — activities where linguistic performance is answerable to worldly coping. When a child learns "hot," the selection pressure is not linguistic — it is the burn. The norms governing human language use are downstream of grounded functions. Language, for humans, inherits its normativity from the embodied predicament it serves.

3.2.3. The application to LLMs

It helps to take the framing just introduced — the world *trains* the organism — and run the analogy in the other direction. If LLM training is relevantly similar to what the frog learns about its environment, the comparison should survive reversal. When we reverse it, however, the disanalogy becomes stark.

Mollo and Millière argue that LLMs satisfy the teleosemantic conditions: their internal states stand in causal-informational relations to worldly features (mediated by training data), and those states were selected (through training) to track those features. We do not dispute that there is genuine selection in LLM training. Gradient descent differentially

retains internal configurations; RLHF further shapes outputs against human-evaluated norms. Shea (2018) has argued that the teleosemantic framework extends beyond natural selection to include learning-based selection, which makes the application to LLMs prima facie plausible. The question is: *selection against what*?

For the frog, the world trains against grounded functions — the organism's need to eat, flee, mate, and navigate. The selection criterion and the represented domain are the same: the world. For the LLM, training selects against linguistic adequacy. Next-token prediction retains states that model textual regularities. RLHF retains outputs that human raters approve of — and yes, raters apply worldly norms like factuality, but they apply them as linguistic judgments about linguistic outputs. The utility criterion that gives teleosemantics its explanatory force — utility for an organism coping with its environment — has been replaced by utility for producing contextually appropriate language.

Human raters are grounded, and their judgments channel worldly norms into the training signal. So, the LLM's selection history *borrows* its world-connection from the raters rather than constituting its own. In biological teleosemantics, this separation never arises: the organism whose states acquire content is the same organism whose survival is a selection pressure. The system that represents and the system that is selected are coupled to the same world through the same body. For LLMs, the system being selected and the system whose grounding constitutes the selection norm are different systems entirely. The LLM is selected to match the *outputs* of a process that was grounded, without itself being grounded.

3.2.4. The scope condition

What emerges is a structural mismatch between teleosemantics and its application to LLMs that operates at three levels. The *selection medium* is language ⟨x⟩ rather than the world: the LLM is trained on text, not on environmental encounters. The *normative source* is borrowed: the

world-connection in the training signal derives from the raters grounding, not from the model's own coupling to what it represents. And the utility criterion is linguistic rather than existential: the function the LLM is selected to perform is language production, not worldly coping.

For organisms, all three levels align automatically: the selection medium is the world, the normative source is the organism's own embodied history, and the utility criterion is survival in its environment. That alignment was so seamless that it never needed stating — it was an invisible precondition of the theory. LLMs are the first systems where the three come apart, and their divergence exposes a scope condition that was always present but never visible: teleosemantics requires that selection occur in a medium that makes contact with the organism whose states carry information about it. LLMs make this condition visible by being the first systems whose selection medium is the linguistic residue of worldly contact rather than worldly contact itself.

Teleosemantics, then, presupposes an already-coupled organism and redescribes that coupling in normative terms. Applied to a system that lacks the coupling, the description still goes through, because the theory was formulated at a level of abstraction that does not distinguish direct environmental coupling from statistical inheritance through text. But the grounding does not come with the description. Teleosemantics works well for systems that are already grounded. It is structurally silent about what makes grounding possible in the first place.

3.3. The isomorphism retreat

Mollo and Millière invoke teleosemantics precisely to go beyond mere structural correspondence — they grant, citing Shea, that isomorphism is too cheap. But under indirect mediation, their conditions reduce to isomorphism plus inherited selection history. So, the question becomes whether structural correspondence, even when causally produced, suffices for grounding.

Interpretability studies show that LLM internal states mirror worldly organisation — Othello-GPT develops broad-state representations, probing studies find vectors tracking spatial relations, colour properties, categorical structure. A reader might concede everything in the previous section and still hold that this correspondence, by itself, suffices for grounding.

I argue it does not. A photograph of a cat preserves the structural relations of the scene with extraordinary fidelity: spatial layout, colour relations, relative sizes, occlusion patterns. The causal chain is impeccable — light from the real cat struck the sensor, was transduced, was stored. The photograph satisfies the causal-informational condition and was produced by a device designed to track worldly features faithfully. Yet no one would say the photograph *refers* to the cat. It is a trace — an imprint that preserves structure without contacting what it preserves. An AI-generated image of a cat, visually indistinguishable from the photograph, makes the point sharper: identical structural correspondence, no causal contact with any cat at all. If isomorphism were sufficient for grounding, the generated image would be grounded in a cat that does not exist.

LLMs are in the position of the generated image, not the photograph. Their internal states mirror worldly structure because they were trained on text produced by grounded speakers — they inherit the structural trace that worldly contact impressed upon language. The correspondence is real, non-accidental, and functionally sustained. But correspondence is not contact. A map drawn from other maps preserves the geography without visiting the terrain.

3.4. What causal chains transmit

Mollo and Millière argue that training data carries causal-informational traces of the world into LLM weights: text about fire was written by people who encountered fire; statistical regularities in that text preserve worldly structure; training encodes that structure into internal states.

I grant all of this. The causal chain is real. What needs examination is what the chain transmits and what kind of causal relation it establishes.

When a grounded speaker writes about fire, she performs a specific operation: she converts her experiential engagement with fire — {FIRE}, the multi-modal, affect-laden, *locus*-bound pattern formed through her fire-encounters — into a linguistic artifact. Those sentences encode ⟨fire⟩: the distributional profile of "fire," its co-occurrences, its inferential relations, its sentential behaviour. This is what language does: it retains the collectively shareable structure and discards the irreducibly first-personal. The causal chain from world to LLM thus passes through a conversion — from {X} to ⟨x⟩ — that strips out the experiential pole. What reaches training data is ⟨fire⟩: the collective precipitate of fire-encounters, not the encounters themselves. The LLM, trained on this data, develops internal states that model the structure of ⟨fire⟩ with impressive fidelity. That structure reflects worldly regularities, because ⟨fire⟩ was produced by grounded beings whose language use was shaped by their fire-encounters.

There is a further difficulty. Teleosemantics characterizes systems whose internal states are causally *coupled* to environmental features through sensory transduction. When the fly crosses the frog's visual field, the fly's presence causally modulates the detector's firing in real-time: remove the fly, the firing stops. This is a detection relation — the internal state covaries with the worldly feature because the feature controls the state. When the LLM acquires internal states that mirror the use of "fire" in the training corpus, it experienced no fire. Worldly structure is present in its weights because it was present in the text — which was produced by grounded speakers whose fire-encounters shaped their linguistic output. The causal ancestry is real: fire-encounters are among the distal causes of the LLM's weights. But causal ancestry is not causal coupling. The LLM's internal states track regularities in *text*, not regularities in *the world that produced the text*. To apply "causal-informational" equally to the frog's fly-detection and the LLM's text-trained states is to treat inherited distal causation as if it were the proximal coupling teleosemantics

describes. That is a substantial philosophical commitment, not a terminological convenience.

A concrete scenario sharpens the distinction. Consider a group of humans settling on Mars, carrying with them an Earth-trained LLM. From the first day, the settlers' grounded functions are under new selection pressure. Gravity is 0.38g — "heavy" and "light" begin to shift their experiential anchoring and, with it, their metaphorical extensions. "Outside" no longer means open air; it means lethal vacuum. The respiratory vocabulary — "breathe," "fresh air," "suffocating" — reorganises around the omnipresent dependence on life support. Weather words lose their old referents and acquire new ones as dust storms replace rain. None of this is deliberate linguistic reform. It is language doing what language has always done: reshaping itself under pressure from the embodied predicament of its speakers. The settlers' language updates because their grounding updates — they are coupled to a world, and when the world changes, the coupling pulls language with it.

The LLM need not update. Its internal geometry encodes the distributional profile of "heavy," "outside," "breathe," and "storm" as shaped by the entire history of Earth-bound human text. It will continue generating "step outside for fresh air" as a suggestion for relaxation. It will associate "light" with ease and "heavy" with burden at Earth-calibrated magnitudes. It will produce weather forecasts structured around precipitation. Not because it tracks the settlers' world, but because it tracks text produced by organisms who were tracking a different world. Only when new text — written by settlers whose encounters with Martian conditions have reshaped their language — enters a future training corpus would the LLM's associations begin to shift. Its states do not covary with the world. They covary with text about a world. When the world changes and the text hasn't yet, the LLM is revealed for what it is: a map of Earth carried to Mars, accurate in structure, connected to nothing underfoot.

3.5. A consequence

One might object that RLHF escapes the ⟨x⟩ register. Human raters reward factuality, penalize hallucination — and "is this factually accurate?" seems to be a question about world-correspondence, not linguistic coherence. But examine what actually flows through the training loop. A rater reads an output — a linguistic artifact — and evaluates it against her own beliefs, which are grounded in her {X} but expressed as a judgment about ⟨x⟩. What enters the model is a scalar reward signal that adjusts the probability of producing similar ⟨x⟩-level outputs. RLHF introduces extra-linguistic *norms on text*. It does not introduce extra-linguistic contact with the world. The grounding lives in the rater. What crosses into the model is a judgment about ⟨x⟩ quality, not about {X} itself. The selection pressure is real, but it selects for text that satisfies grounded evaluators.

Mollo and Milliere's framework cannot, even in principle, distinguish between two systems: one that tracks worldly features through experiential contact — $\mathcal{R}(\{X\}, \langle x \rangle)$ fully realized — and one that tracks the linguistic traces of worldly features with enough fidelity to satisfy any ⟨x⟩-level evaluation. This is not an epistemic limitation — not a matter of needing better interpretability tools or more sophisticated probing studies. It is a methodological consequence of conducting the evaluation in the same register as the system under evaluation.

4. Developmental Irreversibility

Recall the sequence introduced in the previous section:

$$\{X\} \rightarrow \{X, x^o\} \rightarrow \{X, x^o, x^c\} \rightarrow \mathcal{R}(x^o, x^c)$$

Stage 1: Basic grounding. The pre-linguistic experiential pattern. Animals have this. The cat encounters mice; {MOUSE} forms through serial encounters. No language involved. This is the ground floor that the standard debate skips.

Stage 2: Label integration. The child encounters dogs in contexts where "dog" is heard. The word enters the experiential constellation as an embodied element, producing {DOG, dog°}. The label does not represent the experience; it is absorbed into it through co-occurrence in encounter-events.

Stage 3: Dual function. Through increased linguistic practice, the same word acquires concept-mode operation: "dog" can now function within language autonomously — "Dogs are mammals," "Is that a dog?" — as well as in object-mode, pointing at particular encounters. The child has both dog° and dog^{c}.

Stage 4: Reference proper. $\mathcal{R}(x^{o}, x^{c})$ — the mature speaker coordinates between modes. She can use "dog" to talk about dogs (concept-mode) and to pick out this particular dog in front of her (object-mode, anchored in {DOG}). Reference is the coordination between these modes, not a single link from word to world.

The sequence cannot run backward. Concept-mode mastery (Stage 3) presupposes the availability of object-mode (Stage 2), which presupposes the formation of {X} through encounter-events (Stage 1). Grounding must occur in the learning process itself. It cannot be attributed retrospectively based on functional success at a later stage.

LLMs begin and remain at Stage 3. They master concept-mode with extraordinary sophistication — inferential relations, distributional structure, sentential behaviour. Mollo and Milliere's arguments from post-training and pre-training both claim that selection pressures operating at Stage 3 can establish what requires Stages 1–2. PRU says this is structurally impossible. You cannot start with concept-mode mastery and work backward to the experiential anchor.

Mollo and Millière correctly argue that multimodality is neither necessary nor sufficient for grounding. PRU explains why: what matters is not the modality of input but the *mode of learning*. Serial, *locus*-bound, concern-laden encounter-events generate {X}. Massively parallel training over batches of data — whether textual or multimodal — does not. Adding a camera to a parallel-trained system produces sensory data without the developmental structure that turns data into experiential patterns.

LLMs fail to achieve referential grounding not because they lack structured coordinations, but because their coordinations remain strictly $\mathcal{R}(x^c, y^c)$. Teleosemantic selection mediated by text cannot recover the x^o term required for reference proper.

5. Empirical evidence: object-mode vs. concept-mode

The {X} / ⟨x⟩ distinction makes an empirical prediction: if grounded speakers use words differently in object-mode (x^o, anchored in {X}) and concept-mode (x^c, operating within ⟨x⟩), and if these different modes of use leave distributional traces, then training data produced by grounded speakers should encode this difference — and LLMs, trained on that data, should preserve it in their internal geometry.

This prediction is confirmed. Analysis of word-usage manifolds in LLM embedding space reveals a characteristic geometric signature. Sentences using a word in experientially grounded contexts — ostensive, perceptual, deictic (x^o usage) — occupy tighter, more clustered manifold regions. Sentences using the same word in conceptually elaborated contexts — abstract, inferential, generic (x^c usage) — occupy more diffuse, distributed regions. The tight cluster is the "rod"; the diffuse spread is the "caps." The rod-and-caps geometry appears across word categories and across models.

The finding might initially look like evidence *for* Mollo and Millière: LLMs have internal structure that reflects the distinction between experiential and conceptual engagement with the world. But consider what it actually shows. The geometric signature exists because grounded speakers use words differently depending on whether they are in encounter-mode or inference-mode. When a grounded speaker writes "the dog jumped onto the couch" (x^o mode), her word choices, syntax, and co-occurrence patterns differ systematically from when she writes "dogs are social animals" (x^c mode). These distributional differences — generated by the coupling R({X}, ⟨x⟩) in the speaker — leave traces in the

training data. The LLM, trained on this data, encodes the traces faithfully. The signature is a collective residue of subjective groundings, preserved in the ⟨x⟩ patterns that those groundings generated.

The LLM thus preserves even the distinction between experientially anchored and linguistically autonomous usage — it has the map and the map's legend marking which regions came from direct survey — without having done any surveying itself.

6. Conclusion

How do we ground "grounding"? Mollo and Millière's disambiguation of five notions of grounding provided the entry point, and their teleosemantic argument for LLM grounding provided the strongest available case to engage with. Dialectical engagement revealed, I hope, something about the problem itself.

Natural language can produce concepts about grounding but not grounding itself. Correspondence is not contact and causal ancestry is not causal coupling. This is because basic grounding runs from world to organism, not from language to world; that language coordinates with a connection already in place, which I explored via the {X}/⟨x⟩/$\mathcal{R}$ notation.

Teleosemantics earned its credibility on pre-verbal organisms whose selection medium was the world, whose normative source was their own embodied history, and whose utility criterion was existential. LLMs are the first systems where they come apart: the selection medium is text, the normative source is borrowed from human raters, and the utility criterion is linguistic adequacy. This is why Mollo and Millière's argument does not succeed.

Attempting to settle a decade-long debate, I hope to have shown that the grounding debate needs tools that can mark the boundary between what language can transmit and what it cannot — and the discipline to respect that boundary when evaluating systems that operate entirely within the transmissible register (⟨x⟩). I argued PRU is one candidate for such a tool.

As I articulated it, the sequence from basic grounding through label-integration to dual-mode reference is developmentally irreversible. And the rod-and-caps finding provided preliminary empirical evidence that the {X} / ⟨x⟩ distinction leaves measurable geometric traces in embedding spaces, which the LLM faithfully preserves without instantiating the distinction that produced them.

Mollo and Millière ask how vector embeddings can acquire meaning. The question presupposes that meaning is the kind of property an internal item can come to possess. On the Wittgensteinian view developed here, this is already the wrong grammar of the problem. Meaning is not first attached to an item and then connected outward to the world. It is constituted in practice: in language-games, in patterns of use, and in the ongoing coupling between speakers and the world they inhabit.

This is why the Mars settlers language updates. Their words shift because their form of life shifts. "Outside," "air," "weather," "heavy," and "dangerous" are pulled into new patterns by the embodied predicament of the speakers. The practice moves, and language moves with it.

LLMs do not participate in such a practice. They encode the linguistic residue of practices carried out by others and manipulate that residue with remarkable competence. Their internal states may preserve world-traces, including traces of how grounded speakers use words under different conditions. But preserving the residue of a practice is not the same as being a participant in the practice that constitutes meaning.

Mollo and Millière's teleosemantic conditions identify when such residue is well tuned to the world from which it descends. They show, at most, that LLMs inherit structured world-traces through text and training. What they cannot identify is what they were never designed to identify: a system whose language is answerable to the world through its own ongoing form of life.

References

Cojocariu, F. (2026) *Pattern-Recognition Unity (PRU): A Framework Specification; A Meta-Language for Grounding, Reference, and the Experience–Language Interface*. Available at http://dx.doi.org/10.2139/ssrn.6285878

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

Dumitru, M. (2004). Denotare şi descripţie: un criteriu al referinţei pentru termenii singulari. In *Explorări logico-filozofice* (pp. 50-121). Humanitas.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1–3), 335–346. https://doi.org/10.1016/0167-2789(90)90087-6

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

Millikan, R. G. (2017). *Beyond concepts: Unicepts, language, and natural information*. Oxford University Press.

Millikan, R. G. (1989). Biosemantics. *Journal of Philosophy*, 86(6), 281–297. https://doi.org/10.2307/2027123

Mollo, D. C., & Millière, R. (2025). The vector grounding problem. *arXiv preprint arXiv:2304.01481*. https://doi.org/10.48550/arXiv.2304.01481.

Neander, K. (2017). *A mark of the mental: In defense of informational teleosemantics*. Cambridge, Mass.: MIT Press.Neander, K. (1991). Functions as selected effects: The conceptual analyst's defense. *Philosophy of science*, *58*(2), 168–184. https://doi.org/10.1086/289610.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756

Shea, N. (2018). *Representation in cognitive science*. Oxford University Press.

Thompson, E. (2010). *Mind in life: Biology, phenomenology, and the sciences of mind*. Harvard University Press.

---

tipografia.unibuc@unibuc.ro

tel: 0799 210 566

Tiparul s-a executat la Tipografia
Editurii Universităţii din Bucureşti – Bucharest University Press (EUB – BUP)
