Routledge
Taylor & Francis Group

Philosophical Psychology

ISSN: 0951-5089 (Print) 1465-394X (Online) Journal homepage: www.tandfonline.com/journals/cphp20

Merely virtual virtue? The empathy machine hypothesis and the promise of virtual reality

Charles Rathkopf & Jan-Hendrik Heinrichs

To cite this article: Charles Rathkopf & Jan-Hendrik Heinrichs (29 Aug 2025): Merely virtual virtue? The empathy machine hypothesis and the promise of virtual reality, Philosophical Psychology, DOI: 10.1080/09515089.2025.2549085

To link to this article: https://doi.org/10.1080/09515089.2025.2549085

Published online: 29 Aug 2025.

PHILOSOPHICAL PSYCHOLOGY
https://doi.org/10.1080/09515089.2025.2549085

Routledge
Taylor & Francis Group

Merely virtual virtue? The empathy machine hypothesis and the promise of virtual reality

Charles Rathkopf and Jan-Hendrik Heinrichs

INM-7 (Behavior and Brain Sciences), Forschungszentrum Jülich, Jülich, Berlin, Germany

ABSTRACT

Virtual reality (VR) induces a radical psychological reorientation. Yet descriptions of this reorientation are often steeped in theoretically misleading metaphors. We offer a more measured account, grounded in both philosophy and cognitive psychology, and use it to assess the claim that VR promotes moral learning by simulating another's perspective. This hypothesis depends on the assumption that avatar use produces experiences sufficiently similar to those of others to enable empathic growth. We reject that assumption and offer two arguments against it. Empathy relevant to moral learning requires interpretive effort and contextual understanding, not just a shift in perspective. And VR's open-ended, user-driven structure tends to reinforce prior assumptions rather than unsettle them. Still, avatar use may have a different effect on moral learning, which we call self-fragmentation. By loosening the boundaries of the self, VR may expand the range of people one is disposed to empathize with.

ARTICLE HISTORY
Received 21 February 2025
Accepted 12 August 2025

KEYWORDS
Virtual reality; empathy; bias; body schema; virtue

Si on pouvait posséder, saisir et connaître l'autre, il ne serait pas l'autre.

Emmanuel Levinas,

Le temps et l'autre

A platform for moral learning

Virtual reality (VR) is a bewildering technology. When you enter a richly designed virtual world, you undergo a profound psychological reorientation. You adopt a new interpretation not only of your environment, but also of your own body. It is hard to imagine a phenomenon more deserving of philosophical attention. Nevertheless, philosophers have had little to say about it. Moreover, the philosophical work that has been done, although

CONTACT Charles Rathkopf Institute for Neuroscience and Medicine-7 (Brain and Behaviour) c.rathkopf@fz-juelich.de Institute for Neuroscience and Medicine-7 (Brain and Behaviour), Forschungszentrum, Jülich, Jülich 52428, Germany

insightful, has focused on surprisingly traditional philosophical questions. For example, Chalmers (2017) asks whether perceptual experience in a virtual environment is veridical or illusory, and Schwitzgebel (2019) asks whether it provides support for transcendental idealism.

Here we are concerned with the more pragmatic question of whether virtual reality might serve as a platform for a novel kind of moral learning. This question needs refinement. First, our question is not whether moral learning in a virtual environment is possible at all – that much seems obvious. We can imagine a VR classroom, for example, in which you see a graph about the carbon footprint of jet airplanes, and then decide to cut back on flying. Our question is rather about the distinctive potential of virtual environments to influence moral learning. Secondly, the intended meaning of the phrase "moral learning" is virtue-theoretic. From a consequentialist perspective, the acquisition of some purely descriptive information about the lives of people far removed in space and time might count as moral learning. On our intended meaning, by contrast, moral learning is moral character acquisition. It is a matter of developing praiseworthy reactive attitudes and behavioral dispositions toward those with whom one interacts.

Accordingly, we can refine the overarching question that motivates this article: can virtual environments foster moral improvement in ways that no non-virtual environment can? Though this question has a binary form, it is complex and largely empirical. Rather than attempting to settle it definitively, we focus on a flaw in the reasoning behind a widely assumed answer. By bringing this flaw to light and tracing its influence through the literature, we remove a persistent source of philosophical error, clearing the way for a more productive exploration of alternative hypotheses.

The popular idea that we claim is based on flawed reasoning is known as the *empathy machine hypothesis*. Roughly, the empathy machine hypothesis says that the use of an avatar in a virtual environment improves one's capacity for empathy. If the relevant kind of improvement counts as a form of moral learning, and if the learning mechanism cannot be replicated in non-virtual environments, then the empathy machine hypothesis entails an affirmative answer to our overarching question.

Unfortunately, the arguments that have been used to support it rely on the claim that when users come to control an avatar in a virtual environment, they come to know what it is like to be a person whose non-virtual body resembles the avatar. We argue against this latter claim, and thereby undermine existing support for the empathy machine hypothesis. In the final section of the paper, we return to the overarching question about whether virtual environments facilitate moral learning. We show how the space of possibility is constrained

by our critical arguments, and sketch an alternative learning mechanism that not only withstands the critical arguments we put to the defenders of the empathy machine hypothesis, but also remains resilient against recent arguments expressing skepticism about the moral value of empathy itself.

Avatar adaptation

Immersion and interaction

We begin by giving a brief account of the kind of reorientation one undergoes when adopting an avatar in a virtual reality environment.

Chalmers (2017) defines a virtual reality environment as an immersive, interactive, computer-generated environment. According to Chalmers, an immersive environment is one that generates a perceptual experience from a unified perspective within it. This unified perspective is typically achieved by means of a headset outfitted with multiple perceptual technologies. At minimum, these include (i) a flat-panel display for presenting visual information, (ii) a stereoscope for presenting slightly different views of the display to each eye, and (iii) a suite of mechanisms for tracking head position, including a gyroscope, an accelerometer, and a magnetometer. As the head moves, a software program rapidly integrates information from these sensors, and uses the integrated information to update the image on the screen. The update is cleverly designed to make the objects and surfaces in the scene appear to occupy fixed locations. This movement-dependent updating, along with the sense of depth provided by the stereoscope, gives rise to the impression that you are moving about within a virtual environment, rather than looking at it from the outside. Psychologists sometimes refer to this as a sense of presence (Slater & Sanchez-Vives, 2016).

An interactive environment is one in which you can influence what happens. That influence can take many forms. It does not require a virtual body, per se. In the Oculus Rift game, "The Climb," the player exerts control over the virtual world by means of a pair of hands that can move in three dimensions and grasp objects, but which are represented as free-floating, without any connection to a body. We are particularly interested in cases in which the exterior of the body is represented more or less completely, i.e., the user and possible interaction partners are able to see a full body, whatever its shape, although of course this can be done with variable degrees of resolution. In social VR settings, this visible avatar often serves as a public representation of the user, and can become a vehicle for group interaction and identity expression (Moustafa & Steed, 2018).

Sense of ownership

Interestingly, tight coupling does not require that the direction and magnitude of non-virtual movement correspond exactly to those of the virtual movement. It only requires that the mapping between non-virtual and virtual movement be precise, fast, and reliable. When tight coupling breaks down – for instance, due to a poor internet connection causing a time lag – the user is forced to attend to their virtual body explicitly rather than focusing on the objects of bodily action, as they normally would. This experience is disorienting. In the non-virtual world, when we engage in routine activities, our attention is typically directed toward the objects of our actions rather than the specific mechanics of our movements. Thus, tight coupling maintains the fluidity of natural bodily movement in virtual environments.

The second condition necessary to generate the sense of ownership is body schema integration. This occurs when one's virtual body parts become incorporated into an egocentric representation. In the non-virtual world, the body schema is shaped by multiple sensory inputs about the position and movement of the native body. For example, while walking, stretch receptors in the legs continuously relay information about stride shape, while mechanoreceptors in the skin register the air displaced by movement. In a virtual environment, by contrast, the user must rely on a more limited set of signals, some of which inevitably contradict the information provided by these other sources. This discrepancy poses a cognitive and sensorimotor adaptation challenge.

In virtual reality, visual proprioception functions much as it does in experimental setups that generate body illusions. When strolling down a virtual street, the user sees their hands swinging at their sides and their feet landing slightly in front of their body, just as in non-virtual environments. To enhance visual proprioceptive cues, virtual reality environments sometimes include virtual mirrors, allowing users to observe how their virtual body moves in correspondence with their native body. Meanwhile, proprioceptive feedback from the native body remains available through nonvisual channels. The synchrony between perception and action in this process reinforces the integration of the whole virtual body into the user's body schema.

Finally, body schema integration can also be improved by haptic signals synchronized with visual representations of movement. For example, wearing a vest that delivers vibrations to simulate the force generated by virtual objects interacting with the virtual body can further reinforce this integration.

Homuncular flexibility

The fact that people develop a sense of ownership on the basis of sparse, mostly visual body position signals illustrates that the body schema is surprisingly opportunistic: in the absence of the usual plethora of input signals, it will take what it can get. It is also surprisingly malleable. Artificial body position signals can be effective even in cases in which those signals correspond to a body that diverges substantially from the native one. For example, it is possible to learn skillful control over a virtual body with six limbs rather than four. Interestingly, skill acquisition and ownership seem to develop in parallel: as users gain finer control over a divergent avatar, they often begin to experience it as an extension of themselves rather than merely an external tool. However, this process does not simply replace the original sense of bodily self – it complicates it. Users may come to see their virtual and native bodies as partially negotiable rather than fixed. This suggests that ownership is not only an emergent property of sustained, fluent interaction but also a process that can fragment and multiply, challenging the intuitive unity of bodily self-representation.

This brings us to what is, for the purposes of this discussion, the most noteworthy fact about the psychological reorientation virtual reality requires: given our capacity for homuncular flexibility, we can gain a sense of ownership over an avatar that looks nothing like our native body. As far as we know, there is no established term for describing an avatar that does not look like the native body of the user. Let's call any such avatar divergent. All virtual experience is peculiar, but the experience of acting by means of a divergent avatar is especially so. Moreover, the phenomenology of divergent avatar experience is hard to articulate by means of analogy with other, more familiar phenomena.

The distinctive character of avatar use becomes even more pronounced in social VR environments, where avatars are not only controlled from a first-person perspective but also encountered by others as stand-ins for the user. In these settings, users must manage both their control over the avatar and its role as a publicly visible representation of the self. As Freeman and colleagues have shown, users often develop strong identification with their avatars in social VR, and these identifications can persist beyond the VR session (Freeman et al., 2020; Maloney et al., 2021). This social dimension reinforces the idea that avatars are not mere instruments for action, but are entangled with self-presentation and self-conception.

The uniqueness of this experience is both the inspiration behind, and the purported justification for, the idea that virtual reality might serve as a transformative resource for moral learning. Given that no comparable experience is available without the use of virtual reality technology, proponents of virtual reality have, quite sensibly, become interested in whether the

use of a divergent avatar might give rise to unique kinds of knowledge. More carefully expressed, their question is: what can we learn while acting as a divergent avatar that we couldn't learn otherwise? The empathy machine hypothesis, which is the proposal to which we now turn, is a speculative answer to that question.

The empathy machine hypothesis

The term "empathy machine" was coined by the late movie critic, Roger Ebert, in a speech he gave in 2005. Here is what he said.

> For me, the movies are like a machine that generates empathy. If it's a great movie, it lets you understand a little bit more about what it's like to be a different gender, a different race, a different age, a different economic class, a different nationality, a different profession, [to have] different hopes, aspirations, dreams and fears. It helps us to identify with the people who are sharing this journey with us. (Ebert, 2005)

On Ebert's view, great movies help you understand the inner lives of people unlike yourself, and in virtue of that enhanced understanding, you acquire an enhanced capacity to feel concern for such people. Was Ebert right? Is film an empathy machine, in his sense? This is not an easy question. The answer will depend not only on empirical facts about how movies influence people, but also – as we will discuss in some detail below – on what exactly is meant by the term "empathy" (Batson, 2009).

Ebert is certainly on to something. Movies help us appreciate the complexity and variety of human emotion. Ceteris paribus, improving a person's appreciation of emotional complexity and variety should improve their capacity to cultivate concern for people in unfamiliar circumstances. This improved capacity would amount to a positive change in reactive attitudes toward others and thus a case of moral learning as defined above. However, the term "empathy machine" has come to take on a new and more exotic meaning in discussions of virtual reality, and it is this new meaning with which we are primarily concerned.

The phrase "empathy machine" first came to be associated with virtual reality in the work of academic social psychologists who, since the mid-2000's, had been using virtual reality as a medium in which to construct experimental stimuli. Virtual stimuli are typically more immersive than traditional stimuli, and therefore capable of evoking a wider range of psychological processes. Nevertheless, they can be precisely controlled, and therefore retain some of the benefits usually associated with the simple, non-immersive stimuli that cognitive psychologists have used for decades (Bohil et al., 2011). Some research groups, such as that of Mel Slater at the University of Barcelona, and that of Jeremy Bailenson at Stanford, began studying whether virtual reality can influence a person's degree of empathic

concern (Peck et al., 2013; Rosenberg et al., 2013). In their research, the phrase "empathy machine" became connected with the distinctive first-personal feel of using an avatar, and with the consequences of using divergent avatars in particular.

To make this more concrete, consider one particular study by Peck et al. (2013). The driving question behind this study was whether avatar use might influence racial empathy. The study was performed with exclusively light-skinned participants, who were assigned to one of four conditions, defined by avatar type: light-skinned and embodied, dark-skinned and embodied, alien-embodied (in this condition, the avatar was purple), and non-embodied (this condition involved immersion in the same virtual environment, but without an avatar.) Each group was given time to reorient to their virtual circumstances, and were then exposed to a vignette in which virtual characters with different skin tones walked by. Peck et al. measured implicit racial bias before and after, using the racial implicit association test (Greenwald & Banaji, 1995), which measures response time and accuracy on word and face association tasks.[1] The key finding in this study is that participants in the dark-skinned embodied group – those who had a dark skinned avatar – showed a greater reduction in racial bias toward dark-skinned people, compared to those in the other three conditions.

Other variants of this experiment have been done with race, as well as with other socially salient variables, such as age and gender. What we will call the empathy machine hypothesis is the principle that these experiments have in common. Here as an explicit formulation.

The empathy machine hypothesis

When you spend time controlling an avatar that does not resemble your native body, you increase your capacity for empathizing with people whose bodies do resemble that avatar.

The nature of empathy and the subjective similarity hypothesis

In order to evaluate the empathy machine hypothesis, we have to say something about what empathy is and what the defenders of the empathy machine hypothesis take it to be. It is hard to say what empathy is because the term "empathy" is used to refer to more than one psychological phenomenon. Daniel Batson (2009) identified multiple psychological processes that are regularly described by means of the term, some more affective, some more cognitive. In order for the concept of empathy to do the work that the empathy machine hypothesis requires of it, it will have to have both cognitive and experiential features. Jamil Zaki, one of the most prominent psychologists working on empathy, provides

a tripartite model that satisfies this criterion. Moreover, defenders of the empathy machine hypothesis, such as Bailenson and Slater, explicitly appeal to Zaki's work. We therefore use his account of empathy as a way into the discussion.

Zaki's (2014) model of empathy proposes that empathy consists of three interrelated components: experience sharing, mentalizing, and compassion. Experience sharing involves an automatic emotional resonance with another person's state, whereas mentalizing refers to the cognitive capacity to infer another's thoughts and feelings. Compassion, in turn, is the motivation to act in response to another's suffering.

In discussions of the empathy machine hypothesis, the capacity for mentalizing is particularly significant because the kinds of empathy failures that an empathy machine might help overcome arise when we struggle to interpret and respond to the emotions of those who are socially distant from us. Moreover, when discussing these empathy failures, the kind of mentalizing required is sophisticated and fine-grained. Coarse-grained mentalizing judgments – such as whether someone is currently suffering or not – are relatively easy to make even when the empathic target comes from an unfamiliar social group. We can even make judgments about severe suffering in some non-human animals. Where we are more likely to come up short is when we are dealing with subtle, socially modulated mental states. For example, for academic Westerners like us, it would be difficult to accurately judge how someone working in a Japanese corporation feels when their boss tells them they aren't working hard enough. Minimally, we would have to know something about Japanese notions of honor, and of Japanese corporate culture.

Improving this kind of fine-grained mentalizing is not a fast or automatic process. As Zaki stresses, experience sharing alone is insufficient, and in fact, must often be suppressed to allow the interpretive, theory-driven work of mentalizing to operate effectively. This requires rich contextual knowledge, including an understanding of the target's social world, values, and history. Without this, we risk making shallow or mistaken attributions about what others feel, leading to a distorted form of compassion rather than genuine empathy.

This means that any intervention aimed at improving empathy across social gaps must engage with the mentalizing system in a way that allows for fine-grained interpretation of others' emotional states. Mere exposure to an unfamiliar bodily form – such as taking on a divergent avatar – will not, by itself, cultivate the interpretive depth required for understanding.

With this broad account of empathy in place, we now examine the justification for the empathy machine hypothesis. Specifically, we ask: what psychological process could plausibly explain the connection between a person's capacity for empathy and the way their avatar appears? Whatever

the details, the process will have to involve two causal relationships, which can be expressed concisely by means of the following two claims.

(1) Similarity of experience facilitates empathic connection.
(2) Avatar use facilitates similarity of experience.

From these two causal claims, it follows that avatar use facilitates empathic connection, which, of course, is just shorthand for the empathy machine hypothesis. Our criticism of the empathy machine hypothesis is that its presumed justification relies on both of these causal principles, but only the first one is likely to be right.

The claim that similarity of experience facilitates empathic connection is supported by the fact that empathy demands an accurate assessment of what it feels like to be the person at whom your empathic response is directed. As we argued above, in order to successfully forge an empathic connection, you (or some subpersonal mechanism in your head) must correctly interpret how it feels to be your target, given the circumstances they find themselves in.

Plausibly, people whose outer lives are lived under circumstances that are familiar to you will tend to have inner lives that are also familiar, at least to some extent. These people will be easier to interpret, and their subjective experiences will be easier to assess, compared to people you do not know personally or to people who do not share your cultural background.

This supports the first of our two causal claims as at least quite plausible: similarity of experience plausibly does facilitate empathic connection.

The second principle, which we ultimately want to criticize, deserves more detailed discussion. The first step is to clarify what it says. The compact form in which it appears above may obscure the fact that it is intended to describe a relationship that involves at least three entities: the virtual reality user, the avatar, and the person or group that the avatar is designed to resemble. To make this three-way relation explicit, we offer the following expanded version of our second causal principle.

The subjective similarity hypothesis (SSH)

When you take on a divergent avatar designed to be recognized as a member of some socially salient group in which you are *not* normally counted as a member, your subjective experience will be more similar to the typical experience of members of that group than it would have been, had you instead taken on a non-divergent avatar.

Like many tacit assumptions, the SSH becomes harder to defend as soon as it is exposed. What initially felt plausible, especially in the metaphor-rich context of virtual embodiment, now seems to call out for justification.

We will argue that such justification is lacking. But before we do, it is worth asking why the SSH has been so readily, if only tacitly, accepted. One plausible reason is that virtual reality exploits a deeply ingrained tendency toward dualistic thinking. A growing body of developmental and cross-cultural research suggests that dualistic intuitions are a kind of cognitive default (Bloom, 2007; Chudek et al., 2018; Hood et al., 2012). Virtual reality leverages these intuitions by generating compelling illusions – experiences that make it seem as if one is acting through a body markedly different from one's own. This phenomenology can be deeply persuasive, which is precisely what makes it epistemically risky. As a result, VR doesn't merely reflect dualistic intuitions; it amplifies them.

The result is that theorizing in this domain becomes unusually vulnerable to the influence of the false assumption that the self is a detachable perspective rather than an embodied, historically shaped system. This vulnerability is visible in the recurring use of suggestive language throughout the literature: people "transfer" their consciousness into their avatar (Bailenson, 2018, chap. 3, paragraph 9), their body is "replaced" by a virtual one (Slater & Sanchez-Vives, 2016, p. 8), or they are "embodied" in a different body (Kilteni et al., 2013, p. 604). Although the use of these phrases probably does not reflect a theoretical commitment to mind – body dualism, they point to a background picture in which the mind floats free from the constraints of the native body. And that background view, we suspect, makes tacit acceptance of the SSH far more palatable than it ought to be.

Against the subjective-similarity hypothesis

The argument from open-endedness

We have two reasons for thinking that avatar use cannot provide a means by which to gain knowledge of the subjective experience of others.

First, virtual environments that are built to support avatar use are necessarily open-ended. If the user can genuinely use an avatar, and explore the virtual environment within which that avatar is constructed, then the events that unfold in the virtual environment will have to be controlled, at least in part, by the user's choices. This limits the story-telling capacity of the medium and limits its capacity to provide the rich and contextual information in a time frame that approaches the narrated character's presumed experience. Given user-control, virtual reality cannot convey the typical decisions, habits, body language or characteristic utterances of the kinds of people the avatar is taken to represent. Movies, by contrast, can provide such narrative information about the life of a character that is both rich and contextual. This capacity, however, depends directly on the fact that audiences have no choice about the way and speed in which events unfold.

Because individual viewers have no influence over the unfolding of events, their psychological background has comparatively limited influence on the nature of the viewing experience.

These concerns also explain why narrative-driven media, such as literature and film, remain superior tools for moral learning. Unlike VR, where users author their own experiences, a novel or film forces engagement with perspectives that are structured, contextualized, and independent of the reader's or viewer's biases. This is exactly the kind of cognitively effortful work that overcomes empathy gaps. VR's open-endedness, by contrast, ensures that the user's existing assumptions shape their experience, making it unlikely to challenge them. Ironically, then, the very feature that makes VR seem like an ultimate empathy machine – user agency – is the feature that undermines its ability to promote deep, reflective empathy.

This criticism turns a popular thought about virtual reality on its head. Many people have thought that, if movies are a kind of empathy machine, virtual reality must be an empathy machine of unparalleled strength. But this is exactly backwards. If virtual reality is flexible enough to accommodate interesting forms of avatar use, then, precisely for this reason, it will lack the narrative features that make movies so effective, as Roger Ebert suggested they are, at conveying information about the inner lives of others.

The argument from missing mentality

The second reason takes us back to the implicit separability thesis discussed in the previous section. There, we said that the task of describing avatar adaptation seems to amplify our pre-reflective inclination to think of persons as ghostly things that can be transferred into other bodies. If you take the suggestion of body transfer seriously, it becomes tempting to think that divergent avatar use constitutes a technological solution to the problem of having to gather the kind of rich contextual information that, on our view, is necessary to generate knowledge of the subjective experience of others.

In light of our philosophically careful account of avatar adaptation provided in Section 2, however, it is clear that avatar use offers no such work-around solution. Divergent avatar use is a digitally mediated reinterpretation of proprioceptive signals; not a miraculous opportunity to infiltrate a new bodily form. Moreover, and perhaps more importantly, avatars have no minds. Any information available to you about the psychological states of your target, or about the social conditions that modulate their experience, will be information that you bring into the virtual environment from the outside. Moreover, this very fact is easy to overlook. If subjects do overlook it, they may be tempted to interpret the vivacity of their avatar experience as independent confirmation of their beliefs about the experience of others, rather than as a product of those beliefs, as we have

argued it is. This is the idea that is made vivid by the line from Levinas we chose as our epigraph. We may indeed possess (*posséder*) the avatar, but the very fact that we can possess it, or know it in such an intimate way, shows that the avatar is not a genuine other (*l'autre*).[2]

The problem, then, is not just that avatars lack minds, but that VR deprives users of the very cognitive work required to grasp complex social emotions. The kind of fine-grained mentalizing necessary for understanding unfamiliar emotional experiences cannot be achieved passively or through automatic processes. Instead, it requires active engagement with rich contextual narratives, background knowledge, and interpretive effort. Without these, VR experiences risk giving users an illusion of understanding – a false confidence that they have "walked in someone else's shoes" when they have merely inhabited an empty vessel shaped by their own preconceptions. This mirrors the Levinasian critique of possession: the avatar can be owned, but for precisely this reason, it is not a genuine other.

In conclusion, avatar use falls short of providing the rich contextual information necessary for understanding subjective experiences. This undermines the subjective similarity hypothesis and, consequently, the empathy machine hypothesis.

Comparison to Ramirez

Some philosophers have offered independent reasons for skepticism about the empathy machine hypothesis, most notably Ramirez (2017, 2018). Building on the work of Nagel (1974) and Goldie (2011), Ramirez argues that subjective similarity between oneself and another person is impossible to achieve. His argument relies on the idea that attempts to simulate another's experience inevitably distort that experience: in trying to imagine what it is like to be another, we inevitably import our own background psychology into the simulation, thus preventing genuine knowledge of the target's experience. This line of reasoning leads Ramirez to conclude that, although VR may encourage sympathy, it does not dissolve the epistemic barriers that make genuine empathy so elusive.

While Ramirez's argument provides an independent reason to doubt the subjective similarity hypothesis, we take a different approach. Rather than arguing that it is impossible to acquire knowledge of another's experience, we argue that VR lacks the informational structures needed to facilitate this kind of learning. Even if one rejects Ramirez's skepticism, there are strong reasons to doubt that virtual embodiment fosters deep empathy. As we showed in the argument from open-endedness, VR does not provide the structured, contextualized information needed to support fine-grained mentalizing. Moreover, our missing mentality argument shows that any knowledge the VR user gains about an empathic target must already exist

in the user's cognitive framework prior to entering the VR experience. Thus, even if knowledge of others' experiences is possible in principle, VR does not provide a privileged mechanism for acquiring it.

Anticipating an objection

A recent paper by Zahiu et al. (2023) makes a sophisticated case for a claim that sounds very much like the empathy machine hypothesis, and which might, therefore, be thought to contain the ingredients for an objection to our conclusion. They distinguish between bounded and reflective empathy. Bounded empathy is a species of what psychologists call emotional contagion (Batson, 2009). It is fast and involuntary. When we exercise our capacity for automatic empathy, the mental template we use to decide what sorts of suffering are most worthy of our attention does not change. Reflective empathy, on the other hand, is slow and voluntary. This distinction recalls Zaki's division between the experience-sharing and mentalizing components of empathy. Moreover, Zahiu et al. (2023) claim that reflective empathy is more relevant to moral learning than bounded empathy and emphasize, as we have, that understanding the experiences of others requires the acquisition of rich contextual information to ensure that one's beliefs about those experiences are accurate.

However, when they turn to arguing that virtual reality experience is likely to promote empathy, the important distinction between reflective and bounded empathy is left behind. Their argument emphasizes the fact that, in virtual environments, the user is the author of their own actions. In comparison with the process of reading a novel, which they claim is passive, they argue that "VR exploits the primacy of direct personal action." Although we agree that virtual reality is active in just this way, we see no evidence that such agency promotes the specifically reflective kind of empathy emphasized in the earlier part of their discussion.

Since users shape their own avatar experiences, those experiences will largely reflect their preexisting beliefs about what people who resemble their avatar are likely to feel. As our argument from missing mentality showed, however, developing a capacity for reflective empathy requires exposure to evidence that challenges one's assumptions. As far as we can see, nothing about virtual environments makes them distinctively well-suited to providing such exposure.

Self-fragmentation as bias reduction

In the introduction, we posed an overarching question: can virtual environments facilitate the improvement of virtue by some means that cannot be replicated in any non-virtual environment? Thus far,

we have argued against the claim that divergent avatar adaptation fosters empathy. However, even if that claim *were* true, it would not decisively answer the overarching question. To address the overarching question properly, we must consider an entirely different kind of skeptical worry: skepticism about whether empathy is morally virtuous in the first place.

According to Prinz (2011) as well as a number of others who have followed his lead, empathy is a disposition we should actively reject. Paul Bloom, in his (2017) book *Against Empathy: The Case for Rational Compassion*, offers what is now perhaps the most well-known version of the argument. Bloom's criticism involves two essential theses. The first is that empathy, like other interpersonal attitudes, is heavily biased. We are more likely to be empathetic toward people who look like us, or who share our culture, than we are toward those who look different or live far away. Moreover, this bias has political consequences. Empathy will push us to accept preferential treatment of those who are more similar to us, even if that preferential treatment is unfair. This bias, says Bloom, is morally objectionable. Bloom's second thesis is that empathy is unnecessary for moral motivation, especially in light of the fact that we can instead draw on our capacity for compassion, which is a more universal feeling of moral concern for sentient beings, and which, according to Bloom, does not suffer from similarity bias.[3]

Critical responses to Bloom's book tend to push back against the view that empathy is unnecessary for moral motivation, but accept the view that empathy tends to be biased (Bailey, 2022b; Persson & Savulescu, 2018). We agree with this assessment. Empathy does tend to be biased in morally indefensible ways, but may, nevertheless, constitute a necessary source of motivation for at least some kinds of moral behavior. Bailey (2022a) has argued that empathy is necessary for developing what she calls *humane understanding*: the capacity to recognize others' emotions as intelligible. Demonstrating such recognition, says Bailey, is the only helpful kind of response to some kinds of emotional suffering. Furthermore, Persson and Savulescu (2018) argue that compassion itself depends on an initial empathic response. If empathy is indispensable for compassion, then the appropriate moral response to its biases is not to reject empathy altogether, but to mitigate those biases. And this, we now want to suggest, is something with which avatar adaptation might be able to help.

Similarity bias is a tendency to care more for those we perceive as similar to ourselves. That perception of similarity depends not only on how we see others, but also on how we see ourselves. Our suggestion is that divergent avatar use might loosen the rigidity of our self-conception, and thereby facilitate opportunities to forge empathic connections that we would otherwise have missed. Perhaps this self-fragmentation, as we might call it, would

help to de-bias the distribution of targets we deem worthy of empathic response.

Why might avatar adaptation have this fragmenting effect? The answer is not merely that divergent avatars look different to others. In addition, avatar adaption disrupts our sense of self rather directly. Under normal circumstances, our sense of self emerges from the coherence of multiple streams of information: we have a sense of which actions are possible, what sensory feedback would be produced by each action, and the spatial configuration of the body that would result from any particular action. These forms of self-representation are typically synchronized, as cognitive scientists (Vignemont et al., 2021) and philosophers (Milliere, 2020; Zahavi, 2014) have emphasized. In virtual reality, however, this coherence can be disrupted. For example, avatar experimentation allows users to control a third arm, but without the proprioceptive feedback that normally accompanies bodily action. This disassociation from internal signals can extend to the dimensions of one's native body, breaking the usual alignment of perception and action. As a result, what one can do in virtual reality becomes partially untethered from what one can feel, undermining the "immunity to error through misidentification" that typically accompanies bodily self-knowledge (Evans & McDowell, 1982).

This untethering of action and perception is a disunification of the normally unified thing we call the self. We suppose that under these conditions, self-representations will be more malleable, allowing us to represent ourselves as being more similar to other people than we otherwise could have.

The self-fragmentation hypothesis suggests that avatar use might loosen rigid self-conceptions, allowing for a broader sense of identification. While this is a promising mechanism for reducing empathy bias, it does not eliminate the need for structured cognitive engagement with the perspectives of others. Self-expansion without fine-grained mentalizing risks creating shallow identifications that do not translate into real-world moral insight. To really reduce bias, one must acquire the cognitive tools to interpret unfamiliar mental states, which requires more than embodiment – it requires interpretive labor, deep social exposure, and structured engagement with narratives that challenge preconceptions.

It is important to distinguish our hypothesis from the claim that self-fragmentation is itself a form of empathy. Empathy is a relation between persons; it requires some degree of experiential alignment or interpretive fit between the empathizer and the target. Self-fragmentation, by contrast, involves no such interpersonal matching. It is a change in how one represents oneself, not an attunement to another's experience. Our suggestion is that this kind of self-directed disruption may alter the boundary conditions under which empathy becomes possible, by broadening the class of people

perceived as eligible for concern. Whether it has this effect is an empirical question. The connection between self-fragmentation and empathy enhancement is not logical or necessary, but contingent and, at this stage, speculative.

Still, the idea should not strike a philosophical audience as entirely unfamiliar. It echoes Derek Parfit's reflections on personal identity and altruism. Famously, at the end of Reasons and Persons, Parfit writes that after accepting a deflationary view of personal identity, the boundaries between self and other seemed to him less rigid.

> There is still a difference between my life and the lives of other people. But the difference is less. Other people are closer. I am less concerned about the rest of my life, and more concerned about the lives of others. (Parfit, 1984, p. 281)

While Parfit's conclusion was targeted at an abstract, metaphysical conception of personhood, a similar line of reasoning may also apply to a richer culturally embedded conception of personhood. If we loosen our grip on the idea that others are fundamentally different from us because of their cultural, social, or physical traits, the boundaries between "people like me" and "people unlike me" begin to dissolve. Echoing Parfit, we might say that the difference between me and other (kinds of) people is *less*.

Notes

1. There is a lively and substantive controversy around what, exactly, implicit association tests measure, and whether scores on implicit association tests reliably predict behavior outside the psychology lab (Brownstein & Saul, 2016). Recent work suggests that indirect measures such as the implicit association test (IAT) are poor instruments for detecting individual differences, due to poor predictive validity, test-retest reliability, and a lack of conceptual cohesion across tasks (Machery, 2022; Sripada, 2022). If that is true, we should be wary of drawing any conclusions at all from the Peck et al. study.
   Even if implicit association tests are tracking a meaningful mental variable, it remains unclear whether virtual reality influences that variable in a way that demands an explanation of the effect in terms of a mechanism that operates specifically on input from virtual environments.
2. In English, our epigraph reads approximately as follows. "If we could possess, grasp, and know the other, it would not be the other."
3. Bloom uses the term "compassion," to label a universal feeling of moral concern that he thinks most humans have, and that we ought to cultivate. Zaki, however, uses the term "compassion" to refer to the motivational *component* of empathy. This might make it look as if, within Zaki's framework, Bloom's suggestion is logically impossible. But the problem is not nearly so dire. The difference here is primarily terminological. Moreover, the difference has little bearing on our argument, since we are focused on the broader conceptual landscape rather than on potential label disputes.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ORCID

Charles Rathkopf http://orcid.org/0000-0001-8998-2220
Jan-Hendrik Heinrichs http://orcid.org/0000-0002-0560-8483

Author contributions

Charles Rathkopf and Jan-Hendrik Heinrichs contributed equally to the conception, drafting, and revision of the manuscript. Both authors approved the final version.

Ethical approval

This article does not contain any studies with human participants or animals performed by the authors.

References

Bailenson, J. (2018). *Experience on demand: What virtual reality is, how it works, and what it can do*. W. W. Norton & Company.

Bailey, O. (2022a). Empathy and the value of humane understanding. *Philosophy and Phenomenological Research*, *104*(1), 50–65. https://doi.org/10.1111/phpr.12744

Bailey, O. (2022b). Empathy, sensibility, and the Novelist's imagination. In P. Engisch & J. Langkau (Eds.), *The philosophy of fiction: Imagination and cognition* (pp. 218–239). Routledge.

Batson, C. D. (2009). These things called empathy: Eight related but distinct phenomena. In D. Jean & I. William (Eds.), *The social neuroscience of empathy* (pp. 3–15). Social Neuroscience. Boston Review. https://doi.org/10.7551/mitpress/9780262012973.003.0002

Bloom, P. (2007). Religion is natural. *Developmental Science*, *10*(1), 147–151. https://doi.org/10.1111/j.1467-7687.2007.00577.x

Bohil, C. J., Alicea, B., & Biocca, F. A. (2011). Virtual reality in Neuroscience research and therapy. *Nature Reviews Neuroscience*, *12*(12), 752–762. https://doi.org/10.1038/nrn3122

Brownstein, M., & Saul, J. (Eds.). (2016). *Implicit bias and philosophy, volume, 1: Metaphysics and epistemology*. Oxford University Press.

Chalmers, D. J. (2017). The virtual and the real. *Disputatio*, *9*(46), 309–352. https://doi.org/10.1515/disp-2017-0009

Chudek, M., McNamara, R. A., Birch, S., Bloom, P., & Henrich, J. (2018). Do minds switch bodies? Dualist interpretations across ages and societies. *Religion, Brain & Behavior*, *8*(4), 354–368. https://doi.org/10.1080/2153599X.2017.1377757

Ebert, R. (2005). Speech delivered at the dedication of Roger Ebert's plaque outside the Chicago theatre.

Evans, G. (1982). *The varieties of reference*. Oxford University Press. John McDowell (Edited by).

Freeman, G., Zamanifard, S., Maloney, D., & Adkins, A. (2020). My body, my Avatar: How people perceive their avatars in social virtual reality. Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, 1-8. CHI EA '20, New York, NY, USA, Association for Computing Machinery.

Goldie, P. (2011). 17 Anti-Empathy. In A. Coplan & P. Goldie (Eds.), *Empathy: Philosophical and psychological perspectives* (p. 0). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199539956.003.0018

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4. https://doi.org/10.1037/0033-295X.102.1.4

Hood, B., Gjersoe, N. L., & Bloom, P. (2012). Do children think that duplicating the body also duplicates the mind? *Cognition*, *125*(3), 466–474. https://doi.org/10.1016/j.cognition.2012.07.005

Kilteni, K., Bergstrom, I., & Slater, M. (2013). Drumming in immersive virtual reality: The body shapes the way we play. *IEEE Transactions on Visualization and Computer Graphics*, *19*(4), 597–605. https://doi.org/10.1109/TVCG.2013.29

Lévinas, E. (2014). *Le temps et l'autre* (11th ed). Presses Universitaires De France - PUF.

Machery, E. (2022). Anomalies in implicit attitudes research. *Wiley Interdisciplinary Reviews Cognitive Science*, *13*(1), e1569. https://doi.org/10.1002/wcs.1569

Maloney, D., Freeman, G., & Robb, A. (2021). Social virtual reality: Ethical considerations and future directions for an emerging research space. arXiv. https://arxiv.org/abs/2104.05030

Milliere, R. (2020). The varieties of selflessness. *Philosophy and the Mind Sciences*, *1*(I), 1–41. https://doi.org/10.33735/phimisci.2020.I.48

Moustafa, F., & Steed, A. (2018). A longitudinal study of small group interaction in social virtual reality. In Spencer, Stephen N. (Ed), *Proceedings of the 24th ACM symposium on virtual reality software and technology* (pp. 1–10). ACM.

Nagel, T. (1974). What is it like to Be a bat? *The Philosophical Review*, *83*(4), 435–450. https://doi.org/10.2307/2183914

Parfit, D. (1984). *Reasons and persons*. Oxford University Press.

Peck, T. C., Seinfeld, S., Aglioti, S. M., & Slater, M. (2013). Putting yourself in the skin of a black avatar reduces implicit racial bias. *Consciousness and Cognition*, *22*(3), 779–787. https://doi.org/10.1016/j.concog.2013.04.016

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

Prinz, J. (2011). Against empathy. *The Southern Journal of Philosophy*, *49*(s1), 214–233. https://doi.org/10.1111/j.2041-6962.2011.00069.x

Ramirez, E. (2017). Empathy and the limits of thought experiments. *Metaphilosophy*, *48*(4), 504–526. https://doi.org/10.1111/meta.12249

Ramirez, E. (2018). It's dangerous to think virtual reality is an empathy machine. *Aeon: Ethics*, 26. Accessed 19 08 2025. https://aeon.co/ideas/its-dangerous-to-think-virtual-reality-is-an-empathy-machine

Rosenberg, R. S., Baughman, S. L., Bailenson, J. N., & Szolnoki, A. (2013). Virtual superheroes: Using superpowers in virtual reality to encourage prosocial behavior. *PLOS ONE*, *8*(1), e55003. https://doi.org/10.1371/journal.pone.0055003

Schwitzgebel, E. (2019). Kant meets cyberpunk. *Disputatio*, *11*(55). https://doi.org/10.2478/disp-2019-0006

Slater, M., & Sanchez-Vives, M. V. (2016). Enhancing our lives with immersive virtual reality. *Frontiers in Robotics and AI*, *3*, 74. https://doi.org/10.3389/frobt.2016.00074

Sripada, C. (2022). Whether implicit attitudes exist is one question, and whether we can measure individual differences effectively is another. *Wiley Interdisciplinary Reviews Cognitive Science*, *13*(5), e1613. https://doi.org/10.1002/wcs.1613

Vignemont, F. D., Pitron, V., & Alsmith, A. J. T. (2021). What is the body schema? In Y. Ataria, S. Tanaka, & S. Gallagher (Eds.), *Body schema and body image: New directions* (pp. 3–17). Oxford University Press.

Zahavi, D. (2014). *Self and other: Exploring subjectivity, empathy, and shame*. Oxford University Press.

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), 50. https://doi.org/10.1007/s10676-023-09723-9

Zaki, J. (2014). Empathy: A motivated account. *Psychological Bulletin*, *140*(6), 1608–1647. https://doi.org/10.1037/a0037679