What it's like to be a _____:
Why it's (often) unethical to use VR as an empathy nudging tool

Abstract: In this article, we apply the literature on the ethics of choice-architecture (nudges) to the realm of virtual reality (VR) to point out ethical problems with using VR for empathy-based nudging. Specifically, we argue that VR simulations aiming to enhance empathic understanding of others via perspective-taking will almost always be unethical to develop or deploy. We argue that VR-based empathy enhancement not only faces traditional ethical concerns about nudge (autonomy, welfare, transparency), but also a variant of the semantic variance problem that arises for intersectional perspective-taking. VR empathy simulations deceive and manipulate their users about their experiences. Despite their often laudable goals, such simulations confront significant ethical challenges. In light of these goals and challenges, we conclude by proposing that VR designers shift from designing simulations aimed at producing empathic perspective-taking to designing simulations aimed at generating sympathy. These simulations, we claim, can avoid the most serious ethical issues associated with VR nudges, semantic variance, and intersectionality.

Erick Jose Ramirez
Santa Clara University
ejramirez@scu.edu

Miles Elliott
Santa Clara University
miles@mileselliott.com

Per-Erik Milam
University of Gothenburg
per.milam@gmail.com

***Penultimate draft - Accepted for publication in *Ethics and Information Technology*.
Please cite the published version.***

In one form or another, virtual reality (VR) devices have been around since at least the 1960s. In 2016, the Oculus Corporation (a division of Facebook) released the first widely available modern VR device, the Oculus Rift. Later the same year, the Valve Corporation released their own VR hardware, the HTC Vive. These devices represent the first successful generation of modern mass-market VR head-mounted hardware and new, more advanced iterations are already available.

One of the distinctive features of VR is its ability to generate a phenomenon known in the psychological literature as "presence" (Sanchez-Vives & Slater 2005; Cummings & Bailenson 2016). Some researchers have argued that VR can go beyond "presence" and actually give users experiences that can be experienced as being virtually real (i.e., experiences that are treated by the user as if they were real, at least in the moment) (Jacobs & Anderson 2019; Ramirez 2018).

In this article we explore the ethical landscape of what we believe is a largely undertheorized application of VR presence and virtually real experience. In particular, we assess the ethics of using VR for empathy-based nudging. Some VR platforms attempt to use a distinctive kind of VR-based perspective-taking to nudge users toward particular beliefs, attitudes, and actions by eliciting empathetic responses in simulated scenarios. We'll refer to these responses as empathy enhancing experiences (EEEs). Examples include simulations like *1000 Cut Journey* (Cogburn et al. 2018), *Carne y Arena* (Iñárritu 2017), *Becoming Homeless: A Human Experience* (Ogle, Asher, & Bailenson 2018),

and *I am a man* (Ham 2018). These simulations all share a common aim to effect user behavior as a result of exposing them to the perspectives of others via immersive VR.

We argue that, insofar as they aim to nudge their users toward new behaviors by having them experience what it's like to be someone else, these simulations are unethical to produce and to distribute. We do so by appealing to emerging ethical frameworks surrounding technology-based nudges (Hansen & Jespersen 2013; Pinch 2010; Selinger & Whyte 2010; Schubert 2017; Sunstein 2015; Tannenbaum, Fox, & Rogers 2017). In particular, we argue that those who use VR for empathy-based nudging systematically mislead users about the nature of their virtual experiences. It is almost always impossible for VR simulations to show a user "what it's like" from another person's point of view. Such simulations therefore mislead subjects about the nature of their experiences and are at best misleading and at worst amount to an objectionable form of manipulation.[1]

Because users often ground judgments and behaviors on the basis of these false beliefs, we claim that VR empathy nudging of this kind is almost always unethical. Were empathy-based VR nudging the only way to nudge users toward otherwise morally acceptable ends, it may be argued that such nudges are nevertheless morally permissible. However, we show that this is not the case and that alternative, ethically permissible, VR nudges exist.

In the following section, we begin by delineating the specific set of technologies we'll be referring to as virtual reality technologies. Our main concern here is in connecting these technologies with the unique forms of technologically mediated experiences that have come to define them. We'll refer to these as experiences of presence and virtually real experiences. It is these experiences, we claim, that make VR such a potentially powerful vehicle for empathy-based nudging. Following this we'll lay out traditional ethical concerns associated with nudging in order to show that one problem, the problem of semantic variance, looms especially large for VR empathy nudges once we account for the impact of structural intersectionality on experience. We close by offering a partial solution to these ethical issues by showing that designing simulations to enhance sympathetic responses are an ethically superior way of designing VR nudges rather than empathy enhancing simulations that attempt to provide users first-personal access to other minds.

The Many Forms of Virtual Reality

[1] Our claim can also be construed in the following way: empathy-based VR nudgers achieve their nudging effects in virtue of a form of deception that we argue are manipulations of (non-conscious) System 1 and (conscious) System 2 processes. The perceived goods of empathy-based VR nudges almost never outweigh the unethical deception, loss of autonomy, and manipulation associated with VR EEEs especially in the light of available, and less problematic, alternatives to VR empathy-based nudging. Some authors have occasionally framed the ethics of nudges in terms of System 1 and System 2 manipulation (see Sunstein 2015).

The term "virtual reality" was coined by Antonin Artaud to describe a particular kind of theatrical phenomenon (Artaud, 1958; Cogburn & Silcox 2014) in which both actors and spectators engage in a suspension of disbelief in order to give reality to theatrical performances. While *in reality*, there are only people in costumes on a stage, in Artaud's sense, there is a *virtual reality* where Claudius murders Hamlet's father and marries Gertrude. We are interested in a more limited understanding of virtual reality. In particular, we use the term virtual reality (VR) to refer to a small family of technologies that aim to simulate three dimensional environments, place their subjects within those environments, and which serve as the vehicles to deliver immersive experiences.

In this article we focus on VR hardware systems that use dedicated head-mounted devices (HMDs) which have begun to enter the consumer marketplace. Systems like the Oculus Rift, the HTC Vive, and PlaystationVR represent the first generation of consumer VR systems which we argue are capable of giving users the sort of immersive, emotionally engaging, and interactive experiences that can make them highly effective nudge vehicles. Nearly all first-person empathy-based VR nudge vehicles use these dedicated HMD systems.[2]

In what follows, we first explain how VR can generate presence and virtually real experiences. We believe that VR's ability to generate these virtually real experiences is what lies at the heart of their nudging power. Because our critique of empathy-based VR nudging is grounded upon our belief that EEEs systematically *mislead* and *manipulate* users, it's important that we are as clear as possible about the nature of these experiences.

Presence, Immersion, and Virtually Real Experiences

One of the most distinctive features of VR is that many users report feeling like they are actually inside virtual environments instead of wherever they happen to be in reality. Psychologists refer to this kind of experience as an experience of "presence" or of "being present" in the virtual space. These same researchers will also often distinguish the feeling of presence from similar concepts like the feeling of immersion:

> [P]resence in a [virtual environment] is inherently a function of the user's psychology, representing the extent to which an individual experiences the virtual setting as the one in which they are consciously present. On the other hand, immersion can be regarded as a quality

[2] Augmented reality (AR) devices also exist and can be used as nudge vehicles. Instead of simulating an entirely new three dimensional world for users to inhabit, AR devices (like the now defunct Google Glass) overlay information onto our normal experiences (i.e., by providing a floating directional arrow to guide you through an unfamiliar city as you explore it). While empathy enhancing nudges have currently been developed for VR HMDs, what we say in this paper can extend to AR attempts to generate empathy as well (i.e., by changing how you perceive your own, or other people's, racial or gender identity).

> of the system's technology, an objective measure of the extent to which the system presents a vivid virtual environment while shutting out physical reality. (Cummings & Bailenson 2016)

One reason to focus on VR HMDs like the HTC Vive and the Oculus Rift is that they're currently the most successfully immersive VR systems widely available to researchers and the public. To say that these systems are capable of generating the feeling of presence is to say that the hardware of commercial VR systems is capable of producing stereoscopic images and sound while providing convincing haptic feedback, that allows users to feel psychologically present in a simulated space during the course of a virtual experience (Stone 2000; Williams 2014). However, although presence is an important element of what makes VR a powerful vehicle for empathy-based nudging, presence alone does not capture the distinctive way in which VR empathy nudgers like *1000 Cut Journey* are intended to work.

Beyond the feeling of being present in a virtual space, VR simulations, when designed carefully, can generate what we referred to as "virtually real experiences.[3]" Virtually-real experiences are VR experiences that subjects treat *as if* they are real. Although experiences can be more or less virtually real, we argue that virtually real experiences are more likely to occur in simulated environments with high *context-realism* and *perspectival fidelity*.[4] Context-realism refers to the degree to which a simulation's game world--it's physical rules, setting, the appearance and behavior of virtual agents--are programmed to respond as a user would expect them to in the real world. For example, simulations set on the bridge of a starship, far in the future, or those set in distant or fictionalized fantasy pasts are less context-real than simulations set in the present day and in more familiar contexts. A simulation in which the subject can fly, lift heavy weights, or survive unrealistic amounts of bodily damage is less context-real than one where none of these things are possible. Simulations in which non-player characters (NPCs) respond as users expect real humans to are more context-real than simulations where NPCs behave robotically, and so on.

Perspectival fidelity is a way of referring to the structural design features of a simulation that contribute to generating a user's simulated perspective. Given the nature of human perception, a simulation set in the first person is more perspectivally faithful than one that provides its user with a third person perspective of the same world. Similarly, a simulation containing only diegetic music and sound cues will be more perspectivally faithful than one with non-diegetic sound elements (voice-over, a sound track, etc). Because we do not yet live in a world of ubiquitous augmented reality (AR) graphical overlays, a simulation that contains such overlays and meta-information (as many games do) is less perspectivally faithful than a simulation that lacks these features.[5] The conjecture here is that the

---

[3] Citation omitted for review

[4] Citations omitted for review

[5] What we quickly learn from these results is that a simulation's ability to generate virtually real experiences in its users is at least partially a result of the user's psychology and background beliefs. Users who believe in time travel, ghosts, and dragons are more likely to experience simulations containing these elements as more context-real than users without these beliefs (citation omitted for review).

higher a simulation rates on context-realism and perspectival fidelity, the more likely that simulation is to generate a virtually real experience for a particular user.

Virtually real experiences require more than just the feeling of being present in a virtual world. They require that users treat their VR experience as if they were really happening.[6] For example, although VR games may, in virtue of their hardware's immersive capabilities, allow players to feel present in a virtual landscapes (e.g., to feel as if they're on the ramparts protecting the castle gates or flying a starship, going boldly where no one has gone before), it's unlikely that these game experiences are treated as virtually real by those having them. Were that the case, we would expect users to have more traumatic reactions consistent with real life battlefields while playing violent VR games.

Contrast game experiences like these which are high in presence but low in virtually real experience with the sorts of experiences felt by patients undergoing virtual reality therapies for post-traumatic stress disorder (PTSD) (Rizzo et al. 2010) or with the experiences of patients undergoing virtual reality exposure therapy (VRET) to help them overcome specific phobias (Rizzo et al. 2017). These therapies require the creation of simulations that are both context-real and perspectivally faithful so that users are empowered to slowly gain control over their phobias or traumatic triggers in ways that largely mirror traditional real-life exposure therapies and cognitive therapeutic techniques. Importantly, patients undergoing these VR treatments seem to behave *as if* their virtual experiences are really happening to them. This fact about their VR experiences is part of what seems to make such treatments work so successfully when compared to purely hypothetical or imagined therapeutic approaches (Opriş et al. 2012).

Virtually real experiences thus seem to require not only the feeling of presence, of being "in" the virtual world, but also of treating those virtual experiences as more than interesting or enjoyable fictions. They must include the belief that their experiences are *actually* happening, at least in the moment.[7]

A Primer on Nudge Ethics

Our concern is that VR-based nudges mislead and manipulate users by providing experiences that nudge them toward false beliefs about what it's like to be someone else or some other being; these false

[6] That a user treats an experience as virtually real need not indicate that they will continue to act as if that experience were real after-the-fact. What matters, though, is that subjects who are in the middle of the experience will react to it as if it were real. As a result, context-real and perspectival faithful simulations also share some features with context-real and perspectivally faithful dream experiences (Chalmers 2017).

[7] Subjects may, of course, later deny that they "really" thought they were located in a virtual space. One advantage of relying on physiological and behavioral evidence to define virtually real experience is that we can assess the similarity of a subject's responses (arousal, stress, panic, facial expressions, and so on) to their responses to real-world events. A subject whose palms become sweaty while exploring a virtual Grand Canyon, who carefully tiptoes over to the edge, and who refuses to step into the canyon itself could be said, on our view, to be behaving *as if* the experience were real (even if, when verbally prompted, they may say otherwise).

beliefs are then relied upon by users to make future judgments about their own values, policies, and so on. In order to understand this worry, we must explain how nudges influence our choices and the potential moral danger of using such influence.

Nudges are one way of altering an agent's choice architecture. Choice architecture refers to the arrangement of an agent's decision-making context (Hansen & Jespersen 2013). An agent's choices are shaped not only by their desires and values, but also by structural features of their environment. Prominent features include coercive forces (e.g. enforcement of traffic laws) and incentives (e.g. tax deductions for charitable giving). Like coercion and incentives, nudges modify decision environments in order to influence agents toward making a particular choice (e.g. to save for retirement, be an organ donor, or check Facebook more often). However, unlike these other forms of influence, nudges are not coercive (they do not force a user to change their behavior) and they do not change external incentives. Instead, nudges exploit features of choice architecture, including cognitive biases, that do not shape our explicit preferences but have an effect on what we choose; they "steer people in a particular direction while still allowing them to go their own way" (Sunstein 2015). Nudges come in many forms, including reminders (e.g. the sound a car makes when a seatbelt is not fastened), default options (e.g. opting in to organ donation or a retirement plan), and placement/availability (e.g. the layout of different items in a supermarket or the ranking of search results in a list).

While nudges lack the features that many find morally objectionable in coercive laws and burdensome disincentives, they can be morally problematic in their own right. In particular, nudges threaten to subvert three significant moral values: transparency, autonomy, and individual and public welfare (Bovens 2009).[8] First, a nudge is morally problematic if it is contrary to the interest of its target, whether an individual or society more broadly. For example, a supermarket that nudges customers to buy more expensive items rather than identical cheaper items of the same value would be acting contrary to their interests. Second, a nudge is morally problematic if it undermines the autonomy of those it nudges. For example, a casino that exploits a cognitive bias in order to ensure that patrons continue to gamble even when they know they should stop would be undermining their patrons' autonomy (Schüll 2012). Third, a nudge is morally problematic if its effectiveness relies on a lack of transparency.[9] A smoker can understand that the graphic image on a pack of cigarettes discourages them from smoking and does so despite their awareness of such a transparent nudge. However, the need for transparency can make designing an ethical nudge difficult because many nudges require that

[8] According to Sunstein, a person's welfare is benefited when it makes them "better off, *as judged by themselves"* (Sunstein 2015, 429). We wish to remain agnostic about how best to make sense of welfare for the sake of nudging as it seems sensible to claim that public welfare may reasonably render a nudge permissible even if those nudged may not be materially benefitted (as with opt-out organ donation nudgers).

[9] Some might argue that transparency only matters instrumentally, insofar as it enhances autonomy and its absence undermines it. However, even if this is true, understanding whether and how a nudge is transparent or obscure is still helpful in assessing its ethical status.

their operation or mechanism be obscure.[10] For example, if a smartphone user knows that badges and other notifications cause them to spend more time using a social media app, they may disable those notifications.

We can think of transparency, autonomy, and welfare as different moral criteria that must be taken into account, and sometimes balanced against one another, to determine whether a particular nudge is ethical. To say that a nudge is more ethical when it is transparent is to say that transparent nudges do less to undermine individual autonomy than opaque nudges. Most users understand that the visual and audio prompts delivered by a GPS are nudging them toward taking a particular route toward a desired destination and understanding this fact about the nudge does not remove its power as a nudge.[11] The same can be said for speed bumps, images of diseased organs on packs of cigarettes, and so on. Transparency can make designing an ethical nudge difficult, as some nudges only function when the nudgee is unaware that she is being nudged and in such cases the value of transparency and respect for individual autonomy must be carefully balanced against individual and social welfare.

Additionally, it's important to note that many of the VR nudging experiences we discuss in this article are intended to function as *educative nudges*. Educative nudges "attempt to inform people, so that they can make better choices for themselves" (Sunstein 2015). Provided that "the focus is on increasing people's own powers of agency, educative nudges should not be especially controversial on ethical grounds" (Sunstein 2015). Importantly, educative nudges should aim for accuracy in the information conveyed so that they improve, rather than diminish, the autonomy of nudgees and that they do this by delivering truthful information. Because such nudges have specific purposes, aimed at both providing information and improving the nudgee's (or society's) welfare, the ethics of educative nudges requires special assessment.

Because educative nudges derive their force from the information they represent, it's especially important that such nudges are accurate. Nutrition facts, for example, aim to nudge consumers towards healthier food choices. We can imagine an arguably unethical educative nudge that systematically inflates calorie information for unhealthy foods while systematically decreasing the calorie counts of healthier foods. Call such a nudge the "lying calories" nudge. It might be argued that the "lying calories" nudge is transparent (a reasonably careful user would accurately understand that the caloric information printed on their meal is meant to influence their food choices) though it would be a mistake to see this nudge as fully transparent. Users whose choices are influenced by "lying

[10] It's unlikely that transparency fully removes the nudging power of many nudges. For example, hyper-targeted political ads, intended to emotionally nudge users, may retain some of their nudging power even if the nudge's nature is made transparent to subjects. In this sense, that a nudge is transparent (or not) does not fully determine the permissibility of the nudge. A nudge's transparency is an important piece of the puzzle when assessing the ethics of a particular nudge but it may be outweighed by all things considered judgments of the nudge's effects on autonomy and individual/social welfare.

[11] That we can be nudged in this way is one reason why nudge ethicists have called for the development of conditions of competence and trust among nudge developers (Selinger & Whyte 2010).

calories" may be aware of one facet of the nudge (i.e., the caloric data) but not another (i.e., the misrepresentation of that data). Were users to become aware of the true nature of "lying calories" it's doubtful that it would retain its power as a nudge. By falsely representing caloric data, educative nudges like "lying calories" not only deceive users, they also manipulate them. While some deceptive nudges may be, all things considered, ethical to develop we argue that educative nudges are a special case.[12] Educative nudges are unethically manipulative if they rely on deceptive information for their force. This is even more true if non-manipulative alternatives exist.

In recent years, a new nudge vehicle has been gaining popularity: virtual reality experiences which aim to nudge the user towards or away from certain behaviors by placing the user in a particular virtual context. VR gives nudge designers complete control over a virtual space, allowing them tailor virtual environments specifically so that users can experience presence. We now turn our attention to these VR nudge vehicles.

VR For Empathy Enhancement: A Rundown

We have argued that VR simulations can generate ethically significant experiences and referred to these as virtually real experiences. These experiences, coupled with VR's recent commercial popularity, require that we take seriously the use of VR for influencing choice-architecture and behavior. There are two features of VR empathy-enhancing simulations that require special attention.

First, we claim that these simulations, and others like them, represent themselves *as providing* a first-personal experience of what it is like to be someone else; we also give some evidence to support the claim that users themselves leave these simulations believing that they have been given first-personal access to the experiences of the target person, animal, or group. Second, such empathy-based VR experiences affect user behavior. In other words, we argue that those who undergo these VR empathy simulations are often nudged by them into changing their thoughts, beliefs, and behaviors.

A large, and growing, body of research on the psychology of VR embodiment suggests that users tend to identify with the bodies that they virtually inhabit (Won, Bailenson, & Lanier 2015). This remains the case even when those bodies are very different from their own. Other studies suggest that this feeling of merging identity can extend in surprising ways and can have lasting effects on users even after

[12] For example, urinals will sometimes have a fly printed on their surface as a way to nudge users toward aiming at the ideal place on the urinal to reduce splashing. Such a nudge may, arguably, be ethically deceptive in the all things considered sense. Why would this be? First, such a nudge isn't obviously an educative nudge. The printed fly is deceptive in the sense that it might fool a user into believing, at least temporarily, that a real fly is on the urinal but the force of the nudge is derived from its value in providing users somewhere to aim on an otherwise featureless surface. An astute person may realize that the fly is merely printed on the urinal and nonetheless be successfully nudged by it. Second, such a deception (minor and short-lived though it might be) also is likely to align with the user's own goal of avoiding splash. By contrast, "lying calories" goes beyond deception and manipulates users by subverting their autonomy. We say more about deception and manipulation below when discussing the ethics of VR EEEs.

they leave a virtual space (Aardema, O'Connor, Côté, & Taillon 2010). Some psychologists have used results like these to claim that VR simulations can provide their users with the actual experience of what it is like *to be* the sort of person or creature modeled by a simulation. Although we raise significant problems for this presumption, we believe that this assumption is common among designers of VR empathy nudging simulations.

For example, in one study, conducted by Sun Joo Ahn and colleagues (2016), researchers created a simulation where subjects took on the role of a cow being raised for slaughter. During the study, "participants got down on their hands and knees, and embodied a virtual shorthorn cow in a pasture where they saw their cow avatar directly facing them as if looking into a mirror" (404). Ahn et al., in discussing this experiment, make clear that they're operating under the belief that

> [b]ecause IVEs [immersive virtual environments] allow individuals to put themselves inside the virtual body of an animal, they would *directly feel* the threats it is up against and feel *connected* to its plight. For instance, *sharing the experience* of body transfer of oneself to the cow's virtual body would clearly help people *understand how a cow would feel* being raised for its meat. (Ahn et al. 2016)[13]

Though we take issue with this way of understanding VR perspective-taking, we believe that the language Ahn et al. use to describe their experiment is important. Though we have not yet shown *why* the inference from "having a VR experience of being an x" to "knowing what it's like to *be* an x" is mistaken, we wish to show that this inference is common among those who develop and use VR to produce EEEs.

Consider the following additional examples. *1000 Cut Journey* (Cogburn et al. 2018) is a first-person VR simulation that follows the narrative of a Black man at various stages of his life. Designers of this simulation explicitly claim that, in it, "the viewer *becomes* Michael Sterling, a black man, encountering racism as a young child, adolescent, and young adult" ("1000 Cut Journey Premiers"). The simulation's creator, Courtney Cogburn, has informally noted that "...it seems that white people who go through the experience are moved. Tears are not uncommon. They seem to connect more deeply with something they know and understand conceptually but have never felt in quite the way that happens in the VR experience" (Grimsley-Vaz 2018). Cogburn thus believes that White subjects of *1,000 Cut Journey* can, and indeed do, use their new VR experiences to change their thoughts, beliefs,

[13] Emphasis added. We argue that no simulation can show human users what it is like to be a cow or to feel cow feelings or have cow thoughts. In part, we locate part of the problem with VR empathy simulations as resting on a fundamental mistake about the nature of experience that we see repeated throughout the psychological literature on VR embodiment. Though we say more about this later, we think it confusing to parse out how these researchers think human and cow experiences are related and individuated such that we can say that a person's experience, in VR, of walking on all fours in a virtual pasture, is relevantly like an actual cow's experience of walking naturalistically in a field. See (citation omitted for review) and (Nagel 1974) for more on this general problem in the philosophy of mind.

judgments, and values about race and racism because they now believe that they know what it's like to *be* Michael Sterling.

A similar simulation, *I am a man*, promises its users a similar kind of empathic access to the inner life of a simulated person. The simulation's creator claims that "[t]he VR experience allows one to *literally* walk in the shoes of people who fought for freedom and equality during the civil rights era" (Ham 2018).[14] Here too, users play the role of a Black man through various phases of his domestic and political life. Harnessing the power of VR empathy simulations to nudge a user's beliefs, attitudes, and behavior, director Alejandro Iñárritu conceptualized and created his own VR simulation *Carne y Arena* (Iñárritu 2017). Iñárritu's simulation aims to give users the experience of an undocumented immigrant who is being smuggled into the United States from Mexico while also facing the experiences of evading law enforcement and others dangers.

Iñárritu, when speaking of his VR simulation, said that his "...intention was to experiment with VR technology to explore the human condition in an attempt to break the dictatorship of the frame, within which things are just observed, and claim the space to allow the visitor to go through a *direct experience walking in the immigrants' feet, under their skin, and into their hearts*" ("Carne y Arena" 2017).[15] Our claim, however, is that no simulation, no matter how noble its end goal, can deliver the experience of being someone or something else, and that such simulations thus promise to deliver something impossible to their users. Insofar as subjects of these simulations are nudged by their purportedly educative content, we argue that such nudges are unethical.

Educative nudges, as these are intended to be, are ethical to the degree to which they provide accurate information to those being nudged and thus at least part of their wrongness stems from their inability to deliver to users the educative content they promise. To see why these, and other, issues make VR empathy simulations unethical, we need to now be more specific about the problem with EEEs.

Empathy and the Problem of Semantic Variance

To see the issues with VR perspective-taking, we need to say more about what it means to occupy a perspective and why that creates a particular problem (the problem of semantic variance) for EEEs. Psychologists and philosophers speak of perspective-taking predominantly in terms of empathy. Empathy, however, is a multifaceted concept that refers to capacities as diverse as the activity of mirror neurons (Goldman & Jordan 2013), the capacity to predict and explain other minds (Andrews 2008), the capacity to imagine what it would be like to be someone else's shoes (Goldman & Jordan 2013), or even the capacity to imagine what it would be like to *be* someone else (sometimes called "simulation" empathy) (Goldie 2011).

[14] Emphasis added
[15] Emphasis added

For our purposes, it's most appropriate to focus on "in-their-shoes" and "being someone else" forms of empathy because these forms of empathy explicitly focus on perspective-shifting. For example, "in-their-shoes" empathizing has been described in terms of the following process:

> As attempt at perspective-shifting to Bs psychology will have to involve taking on those aspects of Bs characterization that differ from her own whilst at the same time not being conscious of them as such. The reason for this is that the typical role of these dispositions is passive or in the background in the sense that our conscious thoughts and feelings that feature in our deliberations are shaped by, but are not directed towards, these dispositions. (Goldie 2011)

As Goldie makes clear, this form of empathy can be difficult, if not impossible, in most circumstances because agents are likely to differ in terms of their background experiences or in terms of other non-conscious features of experience and thus block successful perspective shifting. For example, internalized racial biases can affect how we consciously perceive the suffering of others (Perez-Gomez 2020; Avenati, Sirigu, & Aglioti 2010) but such biases have their effects, insofar they have them at all, in non-conscious ways. Thus two individuals with different internalized biases will perceive the same situation differently and their first-person perspective will contain different contents, in virtue of that fact. We'll say more about this study in the following sections.

It will thus be impossible to consciously imagine the effects of these features in the imagination. Some have held out hope for this reason that VR might be useful as a vehicle for "in-their-shoes" empathizing:

> Virtual reality simulations can allow us to circumvent the situational version of the background/foreground problem and thus allow subjects to engage in some forms of in-their-shoes empathizing. Having said this, there are limits to the kinds of situational conditions that can be simulated in virtual reality both practically and, especially, ethically. (Ramirez 2017)

Those same authors, however, are also clear about the limitations of VR in this respect:

> Because VR allows us to re-create situational features that affect a subject's judgments, these features allow us to engage in a limited form of in-their-shoes empathizing...VR does not, however, allow us to change a subject's character...VR does not allow a subject to experience what it would be like to be someone else. Subjects will always bring [their own unique] characterological dispositions to an experience. (Ramirez 2017)

VR simulations like *Carne y Arena*, *1,000 Cut Journey*, and *Becoming Homeless* explicitly aim to nudge users by providing EEEs that are designed to model the lives of people very different from those of its intended users. Such simulations thus run into the sort of perspective-taking problem Goldie and others identify. Trevor Pinch (2010), following Evan Selinger and Kyle Whyte (2010), has referred to this difficulty as the problem of semantic variance:

> Semantic variance simply expresses the point known to all interpretive sociologists that meaning depends upon context. The problem of semantic variance with nudges is not just a philosopher's way of posing a counter sample. It goes to the heart of the issue of how people routinely interact with technology and points to the wider social and cultural framework within which technology is embedded. No theory of nudges will be satisfactory without taking on board these sorts of factors. (Pinch 2010)

The meaning of a nudge is, on this view, often not universal and requires that the nudged person be properly situated within the right context in order for the nudge to have its intended effects on the user.[16] Furthermore, such nudges should operate transparently, respect the autonomy of their subjects, and be aimed at improving the subject's welfare or benefit the public good. If the nudge is educative, then the nudge should also deliver accurate information to the subject.

The problem of semantic variance has special force in the case of VR empathy nudges. We argue that identity is intersectional and that the intersectional nature of identity affects users' experiences in a way that impacts how they perceive their environments. If true, then the intersectional nature of individual perspectives is responsible for the problem of semantic variance in the context of VR EEEs. Furthermore, to the degree that users of VR empathy simulators *are* nudged by them, VR EEEs are unethical. In other words, experiential content is not simply *there* in experience but is impacted by myriad, largely subdoxastic and intersectional features, of an individual's life and sense of identity. Because individuals bring their past experiences and identities *to* an experience, and because these help to give that experience its *content*, VR EEEs cannot, in principle, give users access to the experience of "what it's like" to be someone whose identity lies along other intersectional dimensions.

The most problematic ways EEEs nudge their users are thus through deception, via inculcation of false beliefs, and manipulation, because subjects go on to use these false beliefs to alter their behavior. VR EEEs, recall, are marketed as opportunities for users to experience the lives of people (or beings) whose lived experiences will often be very different from the users lived experience. This is something that

[16] Some nudges might avoid the problem of semantic variance altogether because they nudge their users on the basis of physiological or non-cognitive psychological capacities widely shared regardless of a user's context. To the degree that such nudges are cognitively impenetrable (to the degree that a person's thoughts do not affect the power of the nudge), then they would avoid the problem of semantic variance. Images of diseased organs on the cover of cigarette packs, for example, may trigger evolved disgust responses that nudge users to smoke less frequently regardless of their cultural context. Whether or not any nudges, including the one just mentioned, actually fit these criteria is a matter of significant controversy.

they fundamentally cannot provide. For example, in the case of *1000 Cut Journey,* users are likely to believe that what they have experienced in VR bears significant correspondence to the general life experience of (someone like) Michael Sterling. That is, the user will have formed the false belief that they now know something they didn't know before their VR experience: what it is like to be Michael Sterling.[17] If an empathy enhancer delivers this false representation successfully and if a user then goes on to use this information to change their thoughts, feelings, and behaviors (even if in what may appear to be morally laudable ways), then the simulation would have achieved these effects in an ethically problematic way that includes deception and manipulation.

Semantic Variance and Intersectionality

The concept of intersectionality is multi-faceted and scholars within the many intersectional traditions deploy it in different ways and for differing social, philosophical, and political purposes. While we cannot provide a full account or history of intersectionality here, we think it valuable to distinguish between three forms of intersectional analyses to help us make clear exactly why VR simulations which aim to produce EEEs run into the problem of semantic variance.

In a relatively early paper, Kimberlé Crenshaw (1991) usefully distinguished between three kinds of intersectional analyses which she referred to as structural, political, and representational senses of the term. Each sense of intersectionality, while connected to the others, has been stressed by different authors in unique ways and each has taken on an intellectual history of its own over the last thirty years.

For example, scholars focusing on the concept of political intersectionality often use it to help stress the explanatory value that such an analysis can add to an understanding of the way that systems of social and political oppression interact with, and amplify, one another. Crenshaw herself first introduced the term intersectionality in the late 1980s as a way of making better sense of these axes of socio-political oppression and how they intersect. Crenshaw argued that intersectional approaches to oppression helped make salient that which was ignored by traditional "additive" conceptions of oppression (Crenshaw 1989). Scholars working within the tradition of political intersectionality given form by Crenshaw thus argue that

> [i]t is not just that [black women] have quantitatively greater hurdles to overcome than white women or black men, but that the nature of their oppression reflects a distinctive, complex,

[17] It is, of course, possible that a user may actually share many intersectional features with Michael Sterling (they may have been born in a similar era, grown up in similar SES surroundings, and identify as the same gender and race as the character). In those cases, perspective-taking may, indeed, be more likely to deliver truthful simulated content. However, VR empathy simulations are, as we have demonstrated, almost always targeted at populations very much *unlike* those being simulated (undocumented migrants from Mexico, black men, shorthorn cattle, etc).

> and perhaps irreducible combination of sexism, racism, and other structures of oppression, such as classism and heterosexism. (Gasdaglis & Madva 2020)

Intersectionality, understood in this way, is built to resist a purely additive formulation of oppression or of oppressive systems. Someone whose identity lies at the intersections of being "Black" and a "woman" experiences forms of oppression that cannot be understood merely by combining the forms of oppression associated with a person's race and gender on their own.

Closely connected with this political sense of intersectionality is the representational sense of the term. "Representational intersectionality concerns the production of images of women of color drawing on sexist and racist narrative tropes, as well as the ways that critiques of these representations marginalize or reproduce the objectification of women of color" (Carastathis 2014, 307). Examining the nature of representations of identity by isolating individual components (racial representations, gendered representations, classed representation, etc) will miss essential features of how these representations change when they intersect with one another.

Other theorists have chosen to emphasize structural forms of intersectionality to explain and make sense of the effects that intersecting dimensions of a person's identity (expressed in terms of a person's race, ethnicity, class, gender, sexual orientation, etc.) can have on the *content* of an individual's experiences. On this conception of intersectionality,

> Intersectionality...refers to a type or token of experience faced by members of [intersecting identity] categories, as in experiences had by black women that are not entirely explicable by appeal to being black or to being a woman. (Bernstein 2020, 322)

Crenshaw (1991) herself initially framed the concept of "structural intersectionality" in terms of the relationship between a person's identity and a person's *experience*. According to Crenshaw, structural intersectionality, as a methodological tool, can help us focus on

> ...the ways in which the location of women of color at the intersection of race and gender makes [their] actual *experience* of domestic violence, rape, and remedial reform qualitatively different than that of white women. (Crenshaw 1991, 1245)[18]

[18] Emphasis added. Altough she does not use the term "structural intersectionality," consider also Elena Ruíz' (2017) framing of intersectionality in terms of experience: "As a descriptive term, [intersectionality] refers to the ways human identity is shaped by multiple social vectors and overlapping identity categories (such as sex, race, class) that may not be readily visible in single-axies formulations of identity, but which are taken to be integral to robustly capture the multifaceted nature of human experience" (335).

In much the same way in which oppression-centered conceptions of intersectionality resist a purely additive notion of oppression, structural forms of intersectionality resist a purely perspectival conception of experience. What it's like to be you is not given just by what we would get by watching a video feed from the point-of-view of a neutral camera located where your eyes are located coupled with a basic folk psychological theory of perception. Instead, if we want to have a sense of what it's like to be you, that is, if we want to understand the *content* of your percepts as they appear to you, we must factor for the ways that those percepts are shaped, irreducibly, by a combination of personal historical factors including your self identity, conceptual schemas, internalized biases, and upbringing.

Although all of these ways of construing intersectionality have merit, it is this final sense of intersectionality, structural intersectionality, that we appeal to here in order to raise the problem of semantic variance for EEEs. To see the point we aim to make, consider a non-nudging example where structural intersectionality impacts the content of experience.

Empathic contagion refers to what is usually referred to as the non-cognitive activity of sensorimotor mirror neurons to reflect in the experiencing subject the (inferred) experiences of others (Ramirez 2017). For example, witnessing others smile usually activates the same sensorimotor areas in the witness that would be active if the witness herself were smiling. They may, if especially powerful, even cause her to smile in return.

Empathic distress is the name given to instances of empathic contagion in which the suffering of others is shared via these non-cognitive mirror-neuron mechanisms. One of the goals of VR empathy nudgers is thus to trigger these mechanisms in subjects so that they might share, and thus come to better understand, the suffering of others. However, even a relatively primitive non-cognitive capacity like empathic contagion can be influenced by internalized racial bias and affect the content of individual perception (Avenati, Sirigu, & Aglioti 2010).[19] For example, in one study, participants were shown videos of a hand receiving a painful stimulus (a needle prick). Subjects were shown hands belonging to White and Black subjects. Researchers were interested in seeing whether subjects would respond differently (in terms of empathic distress) to hands belonging to members of different racial groups:

> ...the results support the notion that perceiving bodily stimulations on ingroup members leads to an immediate resonance with affective and sensorimotor components of the observed feelings. In contrast, responses to outgroup members' somatic stimulations are less embodied and automatic and likely rely more on slower controlled processing. (Avenati, Sirigu, & Aglioti 2010)

[19] By primitive, we mean only that mirror-neuron empathic structures appear in many mammals, including rats, and thus are likely to have evolved early relative to other more complex cerebral structures found in humans and other apes (Carillo et al. 2019).

When an additional purple/violet hand was included in the experiment, results shed further light on the operation of internalized non-conscious norms on the mechanisms grounding empathic contagion. To their surprise, in this new experimental context researchers discovered that

> ...a clear sensorimotor contagion was found not only in response to the pain of stranger individuals belonging to the same racial group but also in response to the pain of stranger, very unfamiliar but not culturally grouped, individuals (violet models). By contrast, *no* sensorimotor contagion was found in response to the pain of individuals culturally marked as outgroup on the basis of the color of the skin of a nonfacial body part that did not express any specific emotion.[20] (Avenati, Sirigu, & Aglioti 2010)

Our point in discussing this example is that it strongly suggests that internalized racial biases, one of many intersectionally relevant biases which will feature sub-doxastically in a subject's experience of *any* simulation, can strongly affect the operation of non-conscious elements of experience like empathic contagion. Insofar as empathic contagion gives content to a subject's experience during a simulation (something we think very difficult to deny), it serves as an example of the problem of semantic variance for VR EEEs.[21]

Thus, a VR simulation that presents itself to users as giving them access to the first-personal experience of the suffering of a member of an intersectional outgroup member is likely going to fail to provide the user with *the same experience* as an intersectionally ingroup member would have. This is the case even for something as primitive as the operation of empathic contagion on the perception of pain. Given the much larger ambitions of empathy-enhancing VR nudges, such simulations seem destined not only to fail at successfully transferring experiences to users but to fail in ways that are predictable in advance. Although Avenati's experiments focused on racial identity and empathic pain responses, we see little reason to suppose that their findings wouldn't generalize to showing hands classed as "outgroup" on the basis of gender, nationality, sexual orientation, economic class, and so on.[22]

---

[20] Emphasis added

[21] We pause here to note that the problem we're raising for virtual reality simulations is focused specifically on those wishing to use VR as an empathy enhancing nudge. To use VR in such a way *requires* creating a simulation depicting someone's experience with the intention of having that experience be shared with those who are intersectionality different from the person being simulated. VR simulations designed to provide a generic perspective (i.e., a VR simulation of sitting in a stadium during the Olympics) wouldn't run into any of the ethical issues we're raising here for empathy enhancing nudges. Similarly, it's entirely possible for someone to design a VR simulation of their own experiences and intended for their own use. Insofar as the designer and user of a VR simulation like this are similar to one another, then self-empathy of this sort would also be relatively problem free. We say "relevantly" because there is some evidence that we can fail to empathize with our past selves, especially after undergoing transformative experiences (see Levine 1997). For more on how structural intersectionality can impact empathy even across those who identify as members of the same race or gender see Táíwò (2020).

[22] This is, we want to emphasize, an empirical matter. It may turn out that some intersectional features (race, gender, sex, etc) may be more influenced by internalized bias than others (class, nationality). Our point in producing this example is to note that those who will be subject to (or whom subject themselves to) empathy based VR nudge simulations are unlikely,

The Ethics of VR EEEs

Our concern is that VR simulations of the sort described above deceive and manipulate their users. Earlier we described three moral concerns one might have about nudges: transparency, autonomy, and welfare. Our concern with VR EEEs is that these simulations are obscure rather than transparent and that, insofar as they deceive their users, they are manipulative and undermine users' autonomy. Moreover, because VR EEEs are intended as educative nudges, they must meet higher ethical standards than run-of-the-mill nudges used in, say, marketing. Thus, the ability of these simulations not to mislead their subjects given the fact of semantic variance raises serious doubts about using VR EEEs as educative nudge vehicles. The use of VR as an empathy-enhancing nudge vehicle is rendered more problematic if we're right about the fact that it's possible to create VR simulations that can nudge users toward similar desirable goals *without* manipulating users.[23]

Deception is morally objectionable for many reasons. In the present context, VR users are misled about the nature and significance of their experiences. This is problematic because users of VR empathy-enhancing nudges will be led, systematically, to misunderstand their situation and leave them more susceptible to manipulation. Both factors threaten to undermine an agent's autonomy. First, insofar as autonomy requires adequate knowledge or understanding of one's situation, deception undermines autonomy. In the case of VR empathy nudges, users are thus led to misunderstand their own experiences. Second, deception or trickery is a means of manipulation. Deceiving a person is one way to induce a user to have and act on inappropriate beliefs, attitudes, or emotions (Noggle 2020; §2.2).[24] Manipulation of this sort not only undermines but violates its target's autonomy.

Given the likelihood that users will be deceived by VR-based nudges, there are two ways of interpreting what's going wrong in the case of VR EEEs. One possibility is that designers knowingly produce simulations which claim to deliver impossible content (i.e. an experience of "what it's like" to be a member of an intersection of out-groups). On this view, designers are manipulating users into having

---

for these sorts of reasons, to successfully mirror the experience of those whose bodies they virtually inhabit. This is, in one way, to reiterate the concerns we raised earlier about empathic perspective-taking and the importance of subdoxastic features. See also Ruckmann et al. (2015) for a similar study using fMRI to assess empathic contagion.

[23] We intend the arguments in this paper to be as agnostic as possible with respect to first-order moral theory. Nearly all normative systems make room for the value of transparency, respect for autonomy, and individual/social welfare (though for different reasons and in different ways). Our argument relies on the fact that VR empathy-enhancing simulations *unnecessarily* deceive and manipulate users toward (arguably) good ends. The presence of non-manipulative alternatives leaves us comfortable claiming that most first-order moral theories would converge to find this form of nudging unethical given the presence of alternatives. We thank an anonymous reviewer for helping us to clarify this point.

[24] Users can thus be manipulated in several ways. For example, a user can be manipulated toward forming (and/or acting) on inappropriate beliefs (i.e., that a food has substantially more calories than it really does) but they can also be manipulated if they come to form morally benign beliefs through inappropriate/ deceptive means (e.g., convincing a child that Santa Claus does not exist because he was murdered in the 19th century). In the case of VR empathy-enhancing simulations, we believe both forms of manipulation are present. .

particular beliefs and emotions for paternalistic and greater good reasons. Another possibility is that designers are producing VR simulations without understanding how they are misleading their users. On this view, designers are not necessarily engaged in manipulative trickery, but they are nonetheless guilty of a type of culpable wrongdoing.[25]

Suppose that researchers are designing and deploying these VR EEEs with the understanding that they promise their users an impossible experience as a result of the problem of semantic variance. Granting that the purpose of such simulations is beneficial—reducing racial bias, increasing awareness of homelessness, etc.—why might it be nonetheless wrong to design and deploy them?

First, we should be cautious about how we evaluate nudges like these. Unsurprisingly, people tend to find nudges whose goals they approve of to be less problematic than those whose goals they disapprove of, even if the underlying mechanisms by which the nudge operates are the same:

> ...both laypeople and practising policymakers evaluate policy nudges in ways that are coloured by their political preferences. People tend to view nudges as more unethical, coercive and manipulative when illustrated by policy objectives they oppose compared with objectives they support, or when told that such behavioural interventions have been enforced by a policymaker they oppose compared with one they support. (Tannenbaum, Fox, & Rodgers 2017)

This becomes significant if one thinks, as we and most others do, that the moral status of a nudge depends, at least in part, on the structure of the nudge vehicle (Sunstein 2015; Blumenthal-Barby 2013; Hansen & Jespersen 2013). For the purposes of our analysis, we will bracket the often laudable goals of VR EEEs and focus instead on the threat they pose to a users' autonomy. In much the same way as "lying calories" manipulates users and thus is ethically problematic (despite its laudable goal), VR empathy nudges manipulate users into changing their beliefs and behaviors via deception. "Lying calories" becomes more problematic in light of the fact that it's possible to design an "honest calories" nudge that accomplishes many of the same goals without resorting to manipulation.[26]

---

[25] Insofar as designers of VR empathy-enhancing simulations accept even the weakest forms of structural intersectionality (i.e., that internalized concepts can structure the content of experience) then they ought to know that their simulations cannot succeed at the task they've designed them to do.

[26] Another question, outside the scope of our argument here, is whether "lying calories" would be morally acceptable to use if we lacked a non-manipulative alternative. Any response to this issue would require a complex analysis of the value of the social goods that "lying calories" would help realize. Once we account for all of the institutions that would be involved in maintaining the opacity of "lying calories" and the effects of creating and maintaining institutions designed to paternalistically manipulate public beliefs about health, we're skeptical that "lying calories" could be justified, but it isn't beyond the realm of possibility. We thank an anonymous reviewer for pushing us to clarify this aspect of our argument.

We'll argue at the end of the paper that it is possible to produce VR simulations which can preserve the ethically laudable ends of the current crop of VR EEEs while simultaneously avoiding their ethically problematic means.

Researchers who produce simulations with the intent of producing EEEs create simulations which they either know, or ought to know, cannot live up to their aims. Because subjects who experience these simulations are provided with false beliefs about the contents of their experiences and because they then are meant to use those experiences to modify their values, we argue that these subjects are unethically deceived and manipulated by these simulations. To the degree that subjects are treated in this way by the creators of VR simulations, they are treated instrumentally (as means to positive social ends) and without respect for their autonomy (they are manipulated by the experience instead of being reasoned with).[27]

Given the problem of semantic variance with respect to empathy failure for VR EEEs, it may seem as if VR simulations have little to no place in moral enhancement.[28] We believe this conclusion is too hasty. In the final section of the paper, we suggest that altering the design of the simulation, via shifting it from empathy enhancement to sympathy enhancement can sidestep most of the concerns we have outlined.

VR for Sympathy Enhancement

The problem with VR EEEs is that they provide users with false beliefs which are then used to guide future evaluative judgments. As educative nudge vehicles, accurate (truthful) information is paramount to avoid subverting user autonomy. It would be unethical to nudge consumers with a nudge vehicle like "lying calories" even in the service of the laudable goal of encouraging more healthy eating habits. This remains the case even if highly misleading calorie information would more

[27] A corollary issue here is that manipulations of this sort will fail to make their subjects morally better people even if they succeed in changing their behavior. Writing about behavioral moral enhancements of this sort, John Harris (2013) writes that "...moral enhancement, properly so called, must not only make the doing of good or right actions more probable and the doing of bad one's less likely, but must also include the understanding of what constitutes right and wrong action" (172). Because subjects of VR EEEs are using emotionally-laden false beliefs to change their behavior, it's unlikely that they're behaving, as Harris suggests, on the right kinds of reasons when they act.

[28] It might be argued, however, that such nudgers are ultimately permissible. Why would that be? Because becoming more sympathetic with the plight of those experiencing homelessness or with members of other marginalized communities (*1000 Cut Journey*, *Carne y Arena*, etc) would align with the values of the person using the simulation. That is, such nudgers may be used to manipulate oneself into a position that one ultimately desires and endorses upon reflection. The literature on self-deception is interesting, and vast, (Mele 1997; Bortolotti & Mameli 2012; Bermúdez 2000; Lynch 2016) but it seems that such responses miss the point of the critique. In order to properly consent to the use of such simulations, and in order for such nudges to sidestep the ethical issues we raise here, their function must be *transparent* to the user. Our claim here is that such transparency would break the simulation's power as a nudge. If a subject knows, in advance, that the experiences they will have are *not* the experience of "what it's like" to be the person represented by the simulation, then the simulation would cease to function as an empathy-enhancing nudge and thus, by one's own lights, one ought not desire to use the simulation for empathy-enhancing reasons if one cares about the ethics of nudging.

effectively change consumer behavior. Such nudges would be manipulative and deceptive (and could easily backfire). By the same token, VR empathy-enhancing simulations deliver false information to users about the nature of their experiences and this false information nudges them. As a result users are manipulated. It would be one thing for VR empathy simulations to be the only way to nudge users using VR for socially beneficial ends but we think this is false. In this section we show that it's possible to create non-deceptive VR nudges (i.e., sympathy enhancing simulations) and therefore the fact that VR developers and technologists have focused their attention on empathy-enhancing VR content is especially problematic.

Empathy-enhancing simulations, given their necessarily deceptive nature, fail as morally permissible educative nudges. As nudge vehicles more generally, their lack of transparency and manipulative elements cause further issues. However, we believe that all is not lost with respect to using VR for socially desirable purposes. It is possible to design VR simulations that make users more sensitive to the sufferings and injustices faced by others without resorting to misleading claims about empathic perspective-taking.

Writing about the use of VR simulations in museums, Sydney Thatcher discussed the approach that Michael Goldman, director of the US Holocaust Museum, has taken to using VR technology in the museum. His approach, we argue, contains valuable insights for diminishing the ethical issues with VR simulations stemming from semantic variance and for how to better design educative VR nudgers:

> Goldman…has discussed two issues that have come from displaying VR in the Holocaust Museum. Either the visitor minimizes their own experiences, where they think they should not feel bad for themselves, say, because a friend died of cancer, because a Holocaust victim experienced something worse. Or, the visitor over-empathizes with a Holocaust survivor, where they think they know how it feels to be in the Holocaust. To combat these two scenarios Goldman treats visitors as "engaged witnesses" where they recognize the trauma of others without taking that trauma upon themselves. (Thatcher 2019)

In this passage we see Goldman make two claims relevant to our argument. First, he provides more evidence that VR EEEs can mislead their users into believing they have first-hand access to the experiences of what it's like to be the person they embody in a VR experience (e.g., the experience of a Jewish person in a Nazi concentration camp). The problem of semantic variance for VR EEEs is thus not merely a theoretical problem but one already being wrestled with by early adopters of the technology. Second, however, Goldman, and the US Holocaust Museum, have developed an approach that we argue is an ethically sound strategy for sidestepping some of the problems caused by such simulations. Shifting their simulation so that users become "engaged witnesses" *instead* of extermination camp prisoners requires that these VR simulations shift the simulation's aims from providing users with "in-their-shoes" empathy and toward developing the position of a sympathetic

bystander. These simulations aim to enhance sympathy, not empathy. While this shift may not appear significant, we believe it signals an important change in the ethics of VR simulations aimed at changing moral behavior.

Distinguishing empathy from sympathy is controversial and the terms are sometimes used interchangeably. For our purposes, we've defined empathy in connection with empathic contagion and empathic perspective-taking (either "in-their-shoes" or "simulation" examples). Sympathy, however, is less about mirroring the feelings of another. Instead, sympathetic responses require that agents express an attitude of care or concern with respect to their target. For example, if I empathize with someone who has just taken a bad fall, then I will be pained by their experience in the same ways that they are pained. This response might get in the way of helping them if I empathize very strongly, as mirror-touch synesthetes do (Ward & Banissy 2015). A sympathetic helper would have a different response. By expressing an attitude of care or concern for their target, the sympathetic helper would be moved to help the person who has taken a bad fall. They may feel some negative feelings (it's unclear whether sympathy should always be construed as positively valenced) but their experience will differ markedly from the empathetic viewer.

What would it mean to design a simulation to generate sympathy instead of empathy? Minimally, it requires that the simulation be constructed so as to provide the subject with their own point-of-view as opposed to the point-of-view of a different person. What might that entail? Whereas simulations like *1,000 Cut Journey*, *I am a man*, *Becoming Homeless, Carne y Arena*, and others like them explicitly invite subjects to understand their simulated perspective as belonging to another subject (Michael Sterling, a Black civil rights activist, a person experiencing homelessness, an undocumented migrant, etc.) a sympathy simulator would instead follow the model of the VR simulations in use by the American Holocaust Museum. Subjects of those simulations are invited to see the simulations as something *they* are witnessing and hence don't involve problematic perspective-taking.

Additionally, sympathy, understood as the process of *feeling for* another, requires that simulations built around sympathy encourage their subjects to feel for virtual agents within the simulation. This requires that subjects see themselves as witnesses to the relevant events, perhaps even that subjects see themselves as bystanders. For example, Takeshi Moro has created virtual reality simulations which place viewers into cabins that once housed interned Japanese American citizens. During the course of these simulations, subjects are invited to become engaged witnesses, to see and hear from individuals who were actually interned in these camps, as they tell their stories in virtual reality (Hotchkiss 2019). Given the way that the simulations are constructed, subjects in Moro's simulations are unlikely to leave these simulations thinking that they now have first-personal understanding of what it would have been like to be interned in these camps nor are they likely to think that they understand what it was like to have been a Japanese American citizen in the 1940s. The goal of these simulations is to generate sympathy, to help users feel for those who underwent these experiences by hearing from survivors. In the process,

such subjects may also be nudged by the simulation toward changing their thoughts, feelings, or values about the ethics of internment or about US history.

Similarly, performance artist Jordan Wolfson's virtual reality installation "Real violence" is constructed as a sympathy simulator. In this simulation:

> Viewers are directed to a counter, handed noise-cancelling headphones and virtual-reality goggles, and instructed to grip the railing below them. The video begins with a view of clear sky glimpsed between buildings on a wide Manhattan street, as if you're lying supine on the ground…Then a cut, and there, kneeling on a stretch of sidewalk, is a young man in jeans and a red hoodie, an obscure, plaintive expression on his face as he holds your gaze. A man in a gray T-shirt stands over him: the artist. He takes a baseball bat and whacks his victim in the skull, then drops the bat, drags the man by his legs to the center of the sidewalk, and proceeds to bash his face in with a series of stomps and kicks. Blood gushes. The victim grunts and is silent. In the street, indifferent traffic is lined up bumper to bumper. Pedestrians mill around in the far background. The bat has rolled into the gutter; the batterer retrieves it and carries on. (Schwartz 2017)

Wolfson's installation would count as a sympathy simulation on the view we are proposing. First, the viewer, much as in Moro's simulations, is invited to see the experience as something they witness as opposed to offering them an experience of what it was like for someone else to have seen these events. Viewers are thus invited to understand the simulation on their own terms using their own point of view. Given the graphic content of the simulation, it is likely that viewers will come to sympathize with the victims of violent street crime. While Wolfson's simulation is unlikely to have been designed as a vehicle for sympathetic nudging, it may very well have that effect. Our point in providing the example is to show how simulations geared toward generating sympathy will differ in design from simulations aiming at generating (what we have argued is usually impossible) empathic perspective-taking.

Because such simulations invite subjects to bring their own perceptions, histories, background conceptual schemas, and internalized concepts to their experiences, the problem of semantic variance is unlikely to arise in them as they do in empathic simulations. Subjects of a sympathy simulation are not asked to see themselves as a member of an intersectionally different group but instead to bring their own identity to the experience. While some simulations may fail to nudge certain users, this is an issue that any nudge will encounter (e.g., the pinging sound of a seatbelt indicator may fail to convince all riders to buckle up).

Conclusion

Virtual reality technologies hold tremendous promise. Developers are, even now, creating VR simulations that allow their users to visit faraway, even impossible, places or that let them virtually visit with friends and loved ones across large distances almost as if they were sitting in a room with them. Other developers are attempting to harness VR's immersive powers to help us better see and understand one another in the hopes of generating real-world moral progress across a range of important social, environmental, and political issues. We laud all of these goals and genuinely believe that VR technologies can play a role in entertaining us, enriching personal relationships, and in cultivating our capacity to recognize and act on moral reasons.

However, despite their promise, we have argued that VR simulations which aim to generate empathic perspective-taking by providing their users with the experience of what-it's-like to be someone (or something) else are unethical. We have given several reasons. First, such simulations promise the impossible. Given the realities of semantic variance and structural intersectionality, a simulation of what it's like to be a cow at a slaughterhouse will always fail to accurately represent what it claims to represent to its users. The same goes for simulations which aim to represent the inner-lives of intersectionally varied populations. Unless you're already very much like Michael Sterling, *1,000 Cut Journey* cannot give you access to his inner experiences. This problem is morally significant in various ways.

First, one of the primary ethical constraints on educative nudges is that they deliver accurate information to those being nudged. VR empathy simulations in principle *cannot* do this and thus fail as permissible educative nudgers. They are deceptive. Second, nudges in general must take care to nudge their subjects as transparently as possible. Because of the problem of semantic variance with respect to first-person perspectives, empathy nudgers can *only* derive their nudging powers on the basis of their deception and opacity. They are therefore also manipulative. We argue that these features make such nudge vehicles unethical to design or deploy, despite the often good intentions of their creators.

It is better, all things considered, if we use VR simulations to nudge users on more ethically sound grounds. Designing simulations to generate *sympathy* in their subjects instead of first-personal empathy will resolve most of the ethical issues that we have raised in this article. Simulations which aim to affect the user by making them "engaged witnesses" to the plight of others or to environmental destruction or the horrors of factory-farming are, for that reason, morally superior to empathy simulators. This is because they deliver moral intuitions which subjects can treat as their own (as opposed to whatever subject's perspective users believe they're meant to occupy in the simulation). Indeed, we conjecture that users are also more likely to generate longer-lasting and more powerful nudge effects from these sympathy simulations. All the more so if such simulations are perspectivally faithful and context-real so that they generate virtually real experiences.

References

Aardema, F., O'Connor, K., Côté .S, Taillon A. (2010). Virtual reality induces dissociation and lowers sense of presence in objective reality. *Cyberpsychology, Behavior, and Social Networking*, 13 (10), 429-435

Ahn, S. J., Bostick, J., Ogle, E., Nowak, K., McGillicuddy, K., & Bailenson, J. N. (2016). Experiencing nature: Embodying animals in immersive virtual environments increases inclusion of nature in self and involvement with nature. *Journal of Computer-Mediated Communication*, doi:10.1111/jcc4.12173

Ahn, S.J., Tran Le, A.M., Bailenson, J. (2013). The effect of embodied experiences on self-other merging, attitude, and helping behavior. *Media Psychology*, 16, 7–38,

Alejandro G. Iñárritu: CARNE y ARENA (Virtually present, Physically invisible). (2017). Retrieved from https://www.lacma.org/art/exhibition/alejandro-g-inarritu-carne-y-arena-virtually-present-physically-invisible

Andrews, K. (2008). It's in your nature: A pluralistic folk psychology. *Synthese*, 165 (1): 13-29

Artaud, A. (1958). *The theater and its double*. New York: Grove. Trans. Mary Caroline Richards.

Avenanti A., Sirigu A., and Aglioti S.M. (2010). Racial bias reduces empathic sensorimotor resonance with other-race pain. *Current Biology*, 20 (11), 1018-1022

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

Bermúdez, J.L. (2000). Self-deception, intentions and contradictory beliefs. *Analysis*, 60 (4), 309-319

Bernstein, S. (2020). The metaphysics of intersectionality. *Philosophical Studies*, 177 (2), 321-335

Blumenthal-Barby, J.S. (2013). Choice architecture: Improving choice while preserving liberty? In Coons, C. & Weber, M. (Eds.), *Paternalism*. Cambridge University Press

Bortolotti, L. & Mameli, M. (2012). Self-deception, delusion and the boundaries of Folk Psychology. *Humana Mente*, 5 (20), 203-221

Bovens, L. (2009). The ethics of Nudge. In T. Grüne-Yanoff, & S.O. Hansson (Eds.), *Preference Change Approaches from Philosophy, Economics and Psychology*, 207-219. Dordrecht: Springer

Carastathis, A. (2014). The concept of intersectionality in feminist theory. *Philosophy Compass*, 9 (5), 304-314

Carrillo, M., Han, Y., Migliorati, F., Liu, M., Gazzola, V., & Keysers, C. (2019). Emotional mirror neurons in the rat's anterior cingulate cortex. *Current Biology*, 29 (8), 1301-1312

Chalmers, D. (2017). The virtual and the real. *Disputatio*, 9 (46), 309-352

Choi, S., Jung, K., Noh, S.D. (2015). Virtual reality applications in manufacturing industries: Past research, present findings, and future directions. *Concurrent Engineering*, 23 (1), 40-63

Cogburn, J. & Silcox, M. (2014). Against Brain-in-a-Vatism: On the value of virtual reality. *Philosophy and Technology*, 27 (4), 561-579

Crenshaw, K. (1989). Demarginalizing the intersecetion of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory, and antiracist policies. *University of Chicago Legal Form*, 139-167

Crenshaw, K. (1991). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, 43.6, 1241-1299

Cruz-Neira, C., Sandin, D.J., DeFanti, T.A., Kenyon, R.V., Hart, J.C. (1992). The CAVE: Audio visual experience automatic virtual environment. *Communications of the ACM*, 35 (6), 64–72.

Cummings, J., & Bailenson, J. (2016). How immersive is enough? A meta-analysis of the effect of immersive technology on user presence. *Media Psychology*, 19 (2), 272-309

Cogburn, C., Bailenson, J., Ogle, E., Tobin, A., & Nichols, T. (2018). 1,000 cut journey. SIGGRAPH '18 ACM SIGGRAPH, Virtual, Augmented, and Mixed Reality Article No. 1

Fischer, J.M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge: Cambridge University Press

Gasdaglis, K. & Madva, A. 2020. Intersectionality as a regulative ideal. *Ergo: An Open Access Journal of Philosophy*, 6 (4), 1287-1330

Goldie, P. (2011). Anti-empathy. In Coplan, A. & Goldie, P. (Eds.) *Empathy: Philosophical and Psychological Perspectives*. Oxford: Oxford University Press: 318–30

Goldman, A. I., & L. Jordan. (2013). Mindreading by simulation: The roles of imagination and mirroring. In Baron-Cohen, S., Lombardo, M. & Tager-Flusberg, H. (Eds.) *Understanding other minds: Perspectives from developmental social neuroscience*, 3rd ed. Oxford: Oxford University Press: 448–66

Grimsley-Vaz, E. (2018). Creator of '1000 Cut Journey' uses VR to help white liberals understand racism. *Moguldom.com*. Retrieved from https://moguldom.com/152786/creator-of-1000-cut-journey-uses-vr-to-help-white-liberals-understand-racism/

Guttentag, D.A. (2010). Virtual reality: Applications and implications for tourism. *Tourism Management*, 31 (5), 637-651

Ham, D. (2018). *I am a man*. Retrieved from http://iamamanvr.logicgrip.com/

Hansen, P.G., Jespersen, A.M. (2013). Nudge and the manipulation of choice. *European Journal of Risk Regulation*, 3, 3-28

Harris, J. (2013). Ethics is for bad guys! Putting the 'moral' into moral enhancement. *Bioethics*, 27 (1), 169-173

Hotchkiss, S. (Jun 18, 2019). In San Jose's Japantown, contemporary transience takes on historical weight. *KQED*. Retrieved from https://www.kqed.org/arts/13859833/transient-existence-artobjectgallery-san-jose-japantown

Iñárritu, A.G. (2017). *Carney y arena (virtually present, physically invisible*). United States: Fondazione Prada, Legendary Entertainment

Jacobs, O. & Anderson, N. (May 26, 2019). Virtual reality is reality. *Psychology Today*. Retrieved from https://www.psychologytoday.com/us/blog/virtual-reality/201905/virtual-reality-is-reality

Levine, L. (1997). Reconstructing memory for emotions. *Journal of Experimental Psychology*, 126 (2), 165–77

Lynch, K. (2016). Willful ignorance and self-deception. *Philosophical Studies*, 173 (2), 505-523

Mele, A.R. (1997). Real self-deception. *Behavioral and Brain Sciences*, 20 (1), 91-102

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology* 67, 371–378

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83, 435-450

Noggle, R. (2017). Manipulation, salience, and nudges. *Bioethics*, 32(3), 164–170. doi:10.1111/bioe.12421

Ogle, E., Asher, T., & Bailenson, J. (2018). *Becoming Homeless: A Human Experience*. Virtual Human Interaction Laboratory. Retrieved from http://vhil.stanford.edu/becominghomeless/

Parsons, T.D. (2015). Virtual reality for enhanced ecological validity and experimental control in the clinical, affective and social neurosciences. *Frontiers in Human Neuroscience,* 9 (650)

Parsons, T.D., and Rizzo, A.A. (2008). Affective outcomes of virtual reality exposure therapy for anxiety and specific phobias: a meta-analysis. *Journal of Behavior Therapy and Experimental Psychiatry*, 39 (3), 250-261

Perez-Gomez, J. (2020). Verbal microaggressions as hyper-implicatures. *The Journal of Political Philosophy*, https://doi-org.libproxy.scu.edu/10.1111/jopp.12243

Pinch, T. (2010). Comment on "Nudges and cultural variance." *Knowledge, Technology & Policy*, 23 (3-4), 487-490.

Ramirez, E. (2017). Empathy and the limits of thought experiments. *Metaphilosophy*, 48 (4), 504-526

Ramirez, E. (2018). Ecological and ethical issues in virtual reality research: A call for increased scrutiny. *Philosophical Psychology*, 32 (2), 211-233

Rizzo, A., Difede J., Rothbaum B.O., Reger, G., Spitalnick, J., Cukor, J., & Mclay, R. (2010). Development and early evaluation of the Virtual Iraq/Afghanistan exposure therapy system for combat-related PTSD. *Annals of the New York Academy of Sciences*, 1208, 114-125

Rizzo, A., Roy, M.J., Hartholt, A., Costanzo, M., Beth Highland. K.. Jovanovic, T., Norrholm, S.D., Reist. C., Rothbaum. B., Difede, J. (2017). Virtual reality applications for the assessment and treatment of PTSD. In S. V. Bowles & P. T. Bartone (Eds.) *Handbook of Military Psychology* (453-471). Cham, Switzerland: Springer International

Ruckmann, J., Bodden, M., Jansen, A., Kircher, T., Dodel, R., & Rief, W. (2015). How pain empathy depends on ingroup/outgroup decisions: A functional magnet resonance imaging study. *Psychiatry Research: Neuroimaging*, 234 (1), 57-65

Ruíz, E. (2017). Framing intersectionality. In P.C. Taylor, L.M. Alcoff, & L. Anderson (Eds.) *The Routledge companion to philosophy of race* (335-348). New York: Routledge

Sanchez-Vives, M.V., & Slater, M. (2005). From presence to consciousness through virtual reality. *Nature Reviews: Neuroscience*, 6, 332-339

Schubert, C. (2017). Green nudges: Do they work? Are they ethical? *Ecological Economics*, 132, 329-342

Schwartz, A. (March 20, 2017). Confronting the "shocking" virtual-reality artwork at the Whitney Biennial. *New Yorker*. Retrieved from https://www.newyorker.com/culture/cultural-comment/confronting-the-shocking-virtual-reality-artwork-at-the-whitney-biennial

Selinger, E. & Whyte, K. P. (2010). Competence and trust in choice architecture. *Knowledge, Technology & Policy*, 23 (3-4), 461-482.

Sunstein, C. (2015). The ethics of nudging. *Yale Journal of Regulation*, 32 (2), 414-450

Táíwò, O. (August 2020). Being-in-the-Room privilege: Elite capture and epistemic deference. *The Philosopher*, 108 (4). https://www.thephilosopher1923.org/essay-taiwo

Tannenbaum, D., Fox, C.R., & Rogers, T. (2017). On the misplaced politics of behavioral policy interventions. *Nature Human Behavior*, 1, 130

Thatcher, S. (2019). VR and the role it plays in museums. Retrieved from: https://ad-hoc-museum-collective.github.io/GWU-museum-digital-practice-2019/essays/essay-9/

Ward, J., & Banissy, M.J. (2015). Explaining mirror-touch synesthesia. *Cognitive Neuroscience,* 6 (2-3): 118-133

Won, A.S., Bailenson, J., & Lanier, J. (2015). Homuncular flexibility: The human ability to inhabit nonhuman avatars. *Emerging Trends in the Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource*, DOI: 10.1002/9781118900772.etrds0165