Brendan Shea, 2017 (forthcoming), ”The Problem of Evil in Virtual Worlds.” In: Mark Silcox (ed.), *Experience Machines: The Philosophy of Virtual Worlds.* (London: Rowman and Littlefield).

The Problem of Evil in Virtual Worlds

Brendan Shea[1]

In the original Experience Machine (EM) thought experiment, Robert Nozick provides a number of distinct reasons that might explain peoples' intuitive rejection of the chance to plug in (1974, 42–45). For example, he claims that a life lived in the EM deprives an individual of the opportunity to do certain things, to be a certain sort of person, and to genuinely interact with the external world. Choosing to live in the EM, at least on Nozick's account, may even amount to a certain sort of suicide, insofar as it involves giving up the sorts of character traits that constitute one's self-identity. The barrage of pleasant sensations in the EM simply leaves no room to act courageously in the face of danger, to entertain others with a well-told joke, or to demonstrate generosity or compassion in response to the sufferings of others.

As is evidenced by this book, the exact philosophical target of Nozick's EM argument has been widely debated, though it has often been construed as an argument against various forms of hedonism and related theories of human welfare[2]. Whatever its success in this regard, however, I'd like to focus on a somewhat different aspect of life within the EM: it seems to deny one the prospect of a meaningful life, or the possibility of caring about or loving various causes, people, or ideas in the future[3]. More specifically,

[1] I'd like to thank Mark Silcox, Daniel Estrada, and Patrick Taylor Smith for their helpful comments and suggestions during the development of this chapter.

[2] Sumner (1996, 94–99) gives a sympathetic presentation of an EM argument against classical hedonism, though this sort of argument has been subject to considerable criticism in recent years (Kawall 1999; Sober 2000, 37–40; Crisp 2006; Hewitt 2010). Belshaw (2014) provides an excellent overview of interpretations of the EM argument, as well as potential problems with them.

[3] For two recent, influential accounts of the role that *care* and *love* play in human lives, and the ways in which acting from these motives differs from both egoistic self-interest and impartial moral concern, see Frankfurt (2006) and Wolf (2012). Frankfurt provides a subjective account of the importance of love, according to which loving

hooking up to the EM precludes participation in many of the activities that give structure and direction to people's lives: romantic relationships, parenting, friendships, scientific discovery, artistic creation, and so on. Regardless of how much sensory pleasure the EM might bring us, the cost is simply too high: not only must we abandon our current meaning-giving projects, we must abdicate the possibility of taking up any future project of this type, for however long we live in the EM. This may partially account for the widespread intuition that life in the real world is preferable to life within the virtual world of the EM.

The undesirability of life within the EM, however, hardly shows that meaninglessness is a necessary consequence of life within a virtual world. In this chapter, I'll explore the possibility of specifying an EM scenario that avoids this unfortunate consequence by the incorporation of human-like virtual agents worthy of moral concern. I'll argue that, while this scenario would remedy some shortcomings of the original EM, this sort of worldbuilding would be subject to severe ethical restraints. In particular, the fact that this EM would involve the creation of beings subject to suffering and evil would render the user vulnerable to an analogue of the problem of evil familiar from the philosophy of religion. I'll go on to consider the extent to which common theodicies might be adapted to provide moral constraints on this sort of worldbuilding. I will conclude that, while they illuminate certain necessary conditions of any morally justified worldbuilding, they fail to provide sufficient conditions. This suggests that, insofar as we take the creation of virtual agents with moral status to represent a genuine (though perhaps distant) possibility, we have moral obligations to think carefully about the way our design decisions will affect the circumstances in which these agents will find themselves.

something *makes* that thing valuable to the person. Wolf, by contrast, offers a hybrid subjective-objective account, according to which the meaningfulness of human life rests not only a person's loving and caring about certain things, but also the objective value of these things.

1 THE EXPERIENCE MACHINE AS A CHOICE

Nozick's description of the actual EM thought experiment is relatively short. The reader is told that neuroscientists will ensure that she will be given whatever experiences in whatever combination will be most pleasant, and that she won't know about being in the EM while living there, though she'll have a chance to wake up every two years and choose future experiences. Also, she needn't worry about staying unplugged to serve others, since they will also have the ability to hook up to experience machines, if they so choose. Based on these characteristics of the EM, Nozick assumes (plausibly, it seems, based on the reactions of introductory philosophy students) that readers will reject life the EM in the favor of life in real world.

How might one alter the EM scenario to avoid this quick, intuitive rejection? To begin, it is important to recognize that the rejection is, at least purportedly, a choice between two different options: living life in the EM versus living life in a world in which EMs exist, but not hooking up to one. However, while a great deal of attention has been paid to the first, life-in-the-EM alternative, somewhat less attention has been given to the lives of those who choose not to hook up in such a world. Is this a world in which human lives generally resemble ours, or is it one in which they are radically different? In the first case, our intuitive rejection of the EM might carry considerable weight; in the second case, however, there are reasons for exercising much more caution.

Nozick himself has little interest in life outside the EM, and even directs the reader to ignore the question of who will tend the machines if everyone is plugged in. Nevertheless, his description of this world makes the significant posit that everyone who wants to can plug in and that, because of this, the reader need not worry about staying unplugged in order to tend to the needs of others. In this world, it seems, there is literally *no* unavoidable physical or mental suffering; anyone plagued by pain, sickness, depression, or anxiety need merely plug in, and it will all go away. While these people may well have good reasons for refraining from plugging in, this nevertheless represents a significant difference between our world and that of the EM thought experiment. Moreover, there are reasons for thinking the differences from our world are even more pronounced than they might initially appear. For example, in keeping with the spirit of Nozick's scenario, we might also stipulate that EMs can provide medically optimum care, manage their environmental impact, and so on. In the interest of minimizing the impact of potential moral reasons for rejecting the EM, we might go further: let's suppose that the EMs not only deliver maximum benefits to those currently hooked up, but to all future people as well. In this world, there is simply no possibility that the fruits of future scientific research, parenting, or artistic creation, will ever produce outcomes that are hedonically superior to life within the EM. To the extent that one finds meaning in life by pursuing projects that relate to alleviating either one's own suffering or those of others, life in this world may be deeply unsatisfactory, even if it is ultimately better than plugging in[4].

If the world outside the EM is really as boring as described here, why wouldn't one want to plug in? I suspect that the most significant reason concerns Nozick's contention that hooking up to the EM amounts to a sort of suicide. After all, one is submitting not only to a lifetime (or at least a few years) of pleasant experiences, but also a massive forgetting of the fact that one has made this momentous choice, presumably because this knowledge would undercut the pleasure in the EM world*.* If it is to accomplish this task, this forgetting must extend not only to the decision to hook up to the EM, but to any other memory that might cause one to regret lost opportunities. In signing for the EM, then, one must not only renounce the possibility of acting on behalf of the nonexperiential goals one cares about, but also agree

[4] The fact that there is no unnecessary suffering in the world outside the EM may substantially undercut a person's capacity to cultivate what Bernard Williams calls *categorical desires* (1973, 86–88)*,* which provide the grounds for our desiring to continue living at all. For example, while one is still free to do things such as raise children, undertake scholarly inquiry, or create art, these activities are severed from their role in alleviating the unnecessary suffering of the children, developing life-improving technologies, or bringing happiness and emotional fulfillment to one's audience. To the extent that one's categorical desires are directed at bringing about these sorts of outcomes, one may find that one's inability to "make a difference" in the EM world would make it difficult or impossible to sustain these desires.

to have one's brain altered so that one will conveniently forget all of this. This sort of radical change is, as Nozick's remarks suggest, a threat to one's personal identity. Given this, the intuitive rejection of plugging in should not be surprising, even on the supposition that there isn't that much to care about in the world outside the EM.

2 BUILDING A BETTER EXPERIENCE MACHINE

The discussion in the previous section suggests that, while the world outside the EM would likely be deeply unsatisfactory for many people, it would nevertheless remain preferable to plugging in, given the threat that this poses to personal identity. However, this threat to personal identity appears to relate only indirectly to the fact that one lives in a virtual world, as opposed to a real one. Instead, life in the EM endangers one's identity both by precluding one's ability to pursue important desires and by altering one's memory to prevent one from realizing this has been done. These undesirable aspects of the EM scenario might be eliminated if the virtual world in question provided opportunities for users to genuinely care about things inside the EM. So, for example, consider the scorned lovers or unsuccessful scientists who contemplate plugging in to the EM, but who decide that they would rather continue to pursue the (ever-diminishing) chance of *actually having a relationship* or *actually making a discovery* over the alternative of having a *mere experience* of these things.

One way in which one might design a virtual world to remedy this problem is by populating it with human-like AIs with whom the user could have genuine relationships and, more specifically, to whom the agent could owe moral consideration. While the possibility that one might owe moral consideration to AIs has received some philosophical scrutiny in recent years[5], it has long been a staple of science fiction. One

[5] Dennett (1978) provides an early, and quite nuanced, consideration of attributions of *pain* to robots, which he argues are complicated by the incoherency of our current concepts. He concludes by arguing that, given a satisfactory physiological theory of pain, robots could be built to instantiate it, and "thoughtful people would refrain from kicking a robot" (449). In contrast, Torrance (2007) contends that sentience may be necessarily tied to certain features of organic, biological systems. More recently, Anderson (2011) provides a good discussion of the difficulties

might, for example, think of Isaac Asimov's various robots, the androids from Phillip K Dick's "Do Androids Dream of Electric Sheep?" (1968), the robot child of Brian Aldiss's "Supertoys Last All Summer Long" (1969), or the films (*I, Robot; Bladerunner;* and *AI*) based on these stories. More recent examples one might point to include Data (or the holographic Moriarity) from *Star Trek: The Next Generation,* the Cylons of *Battlestar Galactica,* or Samantha from the movie *Her*. These AIs all demonstrate, to varying extents, the sorts of characteristics that have long been thought central to moral status, including the capacity to have interests, to experience pain and pleasure (or the digital analogues of these), and to exercise autonomous choice. They are capable of entering into meaningful relationships with humans, and their well-being is importantly dependent upon how these relationships turn out. They are harmed when humans ignore their interests (or worse yet, actively seek to frustrate them), and are benefitted when humans assist them in various ways.

There is, of course, a long-running debate over whether (science fiction scenarios aside) it is genuinely possible to develop AIs with human-like characteristics such as sentience, consciousness, or autonomy, and it is beyond the scope of this chapter to answer the various objections that been leveled against it. In any case, the role that virtual agents play in the revised thought experiment here might be interpreted in two ways. I will generally assume that the virtual world in question contains agents that really are worthy of moral consideration. If one objects to this scenario, one might instead assume that, while the virtual agents are not genuinely worthy of moral consideration, the potential user of the EM is justified either in believing that they are, or (more weakly) in believing that this is at least a possibility

in determining whether robots meet various traditional criteria for moral status, while Grau (2011) and Bostrom and Yudkowsky (2014) explore ways in which the unique character of machine intelligence might place limits on the way we design and treat AIs. A number of recent authors have also argued that we may owe moral consideration to machines even in the absence of properties such as sentience or autonomy (Floridi 2002; Floridi and Sanders 2004; Neely 2013; Gunkel 2014).

worth taking seriously. In any case, if the argument below is correct, there would be little reason for a potential user who denied this possibility to consider connecting to the EM in the first place.

If designed correctly, a virtual world populated with human-like AIs would offer a potential user several notable advantages over Nozick's original EM. First, plugging in would not require sacrificing the possibility of genuinely *doing* something or *being* a certain sort of person. In fact, this virtual world might actually provide better opportunities than would forgoing plugging in. In the virtual world, unlike the unplugged world, one's artistic, scientific, and personal projects might genuinely serve to prevent avoidable suffering on the part of the AIs that inhabit it. Second, and closely related to this, this revised EM would no longer require that users forget the fact of their plugging in, as was required in Nozick's original scenario. After all, what the user of the original EM had wanted to forget was the knowledge that their experiences weren't genuine, and that nothing in the virtual world was really worthy of concern or care. In the revised virtual world, this is no longer true. Instead, the user has opted to live in a world where she can reasonably expect to care about something or someone, even if she doesn't yet know what this will be.

Many people might still have good reasons to not plug in, of course. In particular, any person with strong pre-existing commitments to people and projects outside the virtual world might well find the prospect of plugging in unattractive, even if they knew that the virtual world would provide opportunities to cultivate alternatives. Nevertheless, there are good reasons for thinking that this sort of virtual world might hold some attractions even if the real world were not as barren of meaning as Nozick's scenario seems to suggest. People's capacities to lead lives that they find meaningful and fulfilling is after all, significantly impacted by numerous factors outside of their control. Along with the obvious challenges posed by lack of resources or ill health (both of which could presumably be addressed with the virtual world), many people find that their desires are frustrated by their inability to make a difference in the world around them. For every successful artist, researcher, athlete, or political or business leader, there are significant numbers of slightly less talented (or less lucky) people who find that their attempts to make meaningful contributions fall short. Importantly, this sort of phenomenon need not merely reflect a morally suspicious desire for increased social status, or for membership among society's elite. Instead, it is result of the fact that people want their projects to succeed, and this success crucially depends not only on their own choices, but on the actions of many others. Arguably, this sense of powerlessness at the root of many people's dissatisfaction with their lives. People repeatedly fail, often through no fault of their own, to establish friendships and romantic relationships, to find careers that allow them to cultivate their talents, or to find receptive audiences for their ideas and artistic contributions.

Given the current state of technology, the best (and perhaps only) solution for people encountering this sort of problem involves abandoning or modifying goals so that they can succeed. For a great many people, however, this solution may be psychologically unrealistic, since it requires abandoning goals and projects that are deeply rooted in both human biology and existing cultural institutions. For an individual in this situation, choosing life in a virtual world where he or she could achieve meaningful goals may well represent an attractive alternative to meaning-deficient real world. After all, in a custom-designed virtual world, one can be reasonably sure that one's capacities, if properly utilized, really can lead to success.

3 VIRTUAL WORLDBUILDING AND THE PROBLEM OF EVIL

In the previous section, I argued that the incorporation of human-like AIs within the virtual world might serve to resolve some significant worries about Nozick's EM. However, the mere incorporation of such agents does not, by itself, provide an adequate reason for users to plug in. After all, if a potential user's reason for dissatisfaction with real world results from a general sense of powerlessness, this would hardly provide reason to plug into a virtual world filled with AIs whose capacities significantly exceeded their own, and who had little to gain from the user's actions. Instead, a potential user would need to be assured that the AIs would be not only worthy of moral concern, but that their capacities would be limited to the extent that it would be within this particular user's ability to genuinely benefit or harm them in significant ways.

This suggests that plugging into the virtual world might be something like playing a video game where the user is the character upon whose choices everything depends, and whose difficulty is precisely calibrated to the user's own abilities. However, the design of this virtual world is complicated by that the fact it contains agents who differ in significant ways from the sorts of monsters and non-player characters (NPCs) that populate existing video games. Some of these differences are merely practical, in that the cognitive complexity of these agents would make it difficult for the designers of the virtual world to ensure that it could genuinely meet the needs and preferences of potential users. For this purposes of this chapter, however, I'll suppose that these difficulties can be overcome.

A much more significant difference between NPCs and the hypothetical virtual agents concerns the fact that the latter, unlike the former, are worthy of moral consideration. This obviously places constraints on what users ought to do while inhabiting the world. Among other things, users should refrain from gratuitously harming virtual agents, and might plausibly have duties to assist them in certain ways. The sorts of behavior encouraged by games such as the *Grand Theft Auto* series would, for example, raise significant moral worries. This suggests that, whatever else these virtual worlds might be good for, they couldn't provide a morally acceptable means by which users could fulfill their immoral desires to harm or exploit people in various ways[6].

[6] Regardless of whether this is morally acceptable, of course, there is a good chance that the virtual agents inhabiting the worlds might be mistreated by the users. For example, recent research suggests that people, especially young people, are prone to engage in aggressive and destructive "bullying" behavior when interacting with service robots in urban environments (Salvini et al. 2010). While it is possible that such behavior would decrease when and if AIs provided evidence that they were worthy of moral consideration, there is every likelihood that these AIs might face maltreatment by significant numbers of humans.

The designers of the virtual world might attempt to circumvent this problem in various ways. So, for example, they might consider designing a world inhabited by masochistic agents who enjoyed suffering the sorts of injuries that careless or malevolent users might inflict upon them. However, it is unclear how widely applicable such a procedure might be. On the one hand, if users were informed about this engineering workaround in advance, this may well cause them to reject life within the virtual world, since this would serve to undercut the very possibility of making a real difference (albeit, a harmful one) that attracted them to the virtual world in the first place. On the other hand, if the engineers were to systematically hide this feature of the world from potential users, it seems doubtful whether the choices of these users to inhabit these worlds would reveal anything about their willingness (or lack thereof) to choose life within the virtual world. Since the present discussion is premised on exploring the nature of worlds that users would choose to inhabit, I'll leave aside discussion of this possibility.

Even supposing that a user's behavior within the virtual world is perfectly acceptable, however, there may be moral problems with the decision to create the world in the first place. So, for example, suppose that a given virtual world is instantiated only when a particular user chooses to plug in, and that the character of the world depends not just on the design of the underlying software program, but also on the preferences of the individual user. The user's decisions, then, play a key role in determining which sorts of virtual agents will come into existence, and what sorts of lives these agents will have. Their lives might be relatively pleasant (if, for example, the user's virtual world involves sitting in a coffee shop discussing philosophy with virtual agents modeled on her favorite historical philosophers), or they might be much less pleasant, if the agents in the virtual world are subject to the endemic violence, deceit, and coercion that characterizes many current video games. However, it seems highly implausible that the lives of the virtual agents will be perfect*,* since the human user's reasons for plugging in require the ability to impact the lives of such agents in meaningful ways. The fact that virtual agents will be subject to at least some sorts of significant suffering and evil thus seems to follow almost inevitably from the very purpose that the virtual world is designed to serve.

The creation of certain sorts of worlds would clearly be morally impermissible. Specifically, it seems plausible that one ought not create a world in which the virtual inhabitants had, on average, lives not worth living. For example, creating worlds where the inhabitants are subject to unending torture seems clearly immoral, even if the human user found life in these worlds to be deeply satisfactory (for example, perhaps it is only by the actions of the user that a few lucky souls could be saved from this awful fate). However, beyond these extreme cases, matters become more difficult. Would it be acceptable, for instance, to create a world filled with highly competitive agents who regularly lose to the user at some game or other, and who suffer from the sort of jealousy and regret that allows the user to bask in her or his victory? These agents might find their lives to be worth living, at least in some minimal sense. Nevertheless, they might find such a situation deeply dispiriting and frustrating. Is the creation of such a world justified by the fact that, were it not for the user, these agents wouldn't have existed in the first place? Or did the user have some moral obligation to avoid creating a world that so was deeply "unfair"?

The question of what, if any, moral constraints might be placed on the creation of virtual worlds has several close analogues within contemporary philosophy. First, it has connections with the question of whether it is wrong for people to choose to reproduce when they have reasons to believe that the children resulting from these decisions will be less well-off than the children that might have resulted from other decisions[7]. Second, it bears a striking resemblance to the question of whether the omniscient,

[7] The creation of virtual agents who suffer has connections with the so-called "non-identity problem" usually attributed to Parfit (1976; 1984), which involves determining moral culpability for bringing beings into existence whose lives are foreseeably flawed. I will generally assume that one is not obligated to create new beings (even if these beings would be happy), and that, conversely, there are at least some cases where the creation of a new being is morally permissible. Both of the assumptions have been challenged. For example, Rachels (1998) argues that it is (morally) good to bring beings into the world whose lives are generally worthwhile, while Benatar (1997; 2006) argues that that, since this condition can never be met, there is a moral obligation to refrain from procreation.

omnipotent, omnibenevolent God of classical theism could be morally justified in creating a world—such as the one we currently inhabit—with widespread cases of apparently undeserved suffering[8]. In the remainder of this essay, I'll be examining this second case in some detail to consider what, if anything, the debate over the problem of evil might show about the morality of creating virtual worlds inhabited by beings to whom one owes moral concern.

There are, of course, important differences between the human creators of virtual worlds and the divine creator targeted by problem of evil arguments. First, the human creators' incomplete knowledge and limited power may well make it impossible for them to design a world that maximizes the well-being of the potential virtual inhabitants, even if they desired to so. Within this in mind, it may be inappropriate to demand that a given virtual world be the best possible world. However, this deficit of knowledge or power does not alleviate human creators of any moral responsibility for a given world's shortcomings, so long as the creators were capable of making comparative judgements as to the degree and type of suffering likely to be present in the various alternative worlds they might create. So, for example, the human creators may be incapable of predicting with any precision the lives of individual virtual agents. Nevertheless, they could confidently predict that the average inhabitant of a zombie-apocalypse virtual world would be worse off than the average inhabitant of resource-rich world devoted to artistic and scientific pursuits. Second, there is no assumption that the human creators are omnibenevolent, or that the users or creators of these worlds are the proper object of worship by the virtual agents inhabiting the world. This suggests that the human creators need not be held to the same high standard as the divine creator discussed in traditional theodicy. In the latter case, the goal is to establish what sorts of worlds

[8] Mackie (1955) famously argues that the existence of evil is logically incompatible with the existence of God, though this has been influentially challenged by Plantinga (1974, 12–55). Most recent debates on the problem of evil have tended to focus on *inductive* or *evidential* formulations, such as those offered by Rowe (1979; 1991; 1998; 2006). For overviews of the current debate, see Pereboom (2005) and Tooley (2015). The dilemma facing human creators of virtual worlds is, of course, importantly different from that faced by theists. In the former case, the creators want to know which sorts of worlds it might be morally permissible to create; in the latter case, theists seek to show that the actual world is among the morally permissible ones.

might be created by a morally perfect being, of the sort that might plausibly be worthy of worship by the human inhabitants of this world. In the former case, the goal is much more modest: we want to know which worlds a human might create without being morally blameworthy.

With these caveats just mentioned in mind, one can now formulate the problem of evil for virtual worlds in a more precise form.

1. Any virtual world that is worth creating (from the potential user's perspective) will contain significant amounts of evil [suffering, frustration of desire, etc.] by virtual agents.
2. It is morally wrong to create a virtual world with significant amounts of evil, when one has the power and knowledge to create one with lesser amounts of evil.
3. It is within the user's power and knowledge to create a world that contains lesser amounts of evil.
4. So, it is morally wrong to create any virtual world that is, from the user's perspective, worth creating.

Is there any way out of this dilemma—to create a world that contains adequate evil to meet the needs of the human user without violating the demands of morality? In the remainder of the chapter, I will examine the possibility of justifying the inclusion of evil within a virtual world by looking to components of traditional theodicies. These potential justifications will, in effect, amount to objections to (2) and (3) above. The goal here will be to establish with more precision *which sorts* of evil ought to be especially concerning creators of virtual worlds, and what sorts of constraints this might place on the creation of virtual worlds.

4 THEODICIES FOR VIRTUAL WORLDS?

What, if anything, might serve to justify the intentional creation of a virtual world with significant amounts of evil? In the context of the religious problem of evil, various theodicies provide purported answers to this, and seek to identify possible or plausible reasons that God may have for allowing the sorts of evil we see around us. In this section, I'll take a look at a number of the key components of theodicies[9], and consider what relevance, if any, they might have for the creation of virtual worlds.

4.1 Free Will

Free will forms a central element of most theodicies. Specifically, many theists have claimed that libertarian free will is an important good for human beings, and that one necessary consequence of creating a world with free will is the existence of the evil caused by its exercise. Considerations of free will are generally taken to be most directly relevant to explaining the existence of *moral evil,* or the evil inflicted by human beings, as opposed to *natural evil,* or the evil resulting from natural causes such as earthquakes, tsunamis, or disease.

Before considering the applicability of this argument to the case of virtual worlds, it is important to recognize that the conception of free will relevant to virtual worlds is of a different character than that appealed to in the context of theodicies. First, we can set aside contentious claims about the coherence of libertarian free will. After all, it is unclear whether the virtual agents could possess it and, even if they could possess it, there is no evident mechanism by which the human creators of the virtual worlds could grant or deny it to them. Instead, the sort of free will relevant to virtual agents is something closer to a capacity for autonomy, or self-governed action. If this is the case, then important aspects of many traditional free will defenses are not readily applicable to the creation of virtual worlds. In particular, there seems no reason to suppose that the existence of these sorts of autonomous agents logically (or

[9] In this chapter, I make no distinction between responses to the problem of evil that take the form of *theodicies,* and those that take the form of *defenses.* On van Inwagen's (2008) characterization, a theodicy "is an attempt to state the real truth of the matter…about why a just God allows evil to exist" (6), while defenses are stories that "may or may not be true" (7) but which have some other desirable attribute, such as providing reasons that would have justified God in allowing evil. When discussing of the creation of virtual worlds, I will generally assume that we already know the general sorts of reasons for which the user in question created the world, and that our moral assessment of the world-creation can and should take this into account.

metaphysically) entails the existence of evil, in the way that libertarian free will might. Because of these differences, it is not open to the human creator of the virtual world to escape moral responsibility for a particular instance of evil on bare grounds that this evil was the result of a “free action” by a virtual agent. This is not to say, of course, that the virtual agent might not also bear moral responsibility; the point is merely that this has no bearing on the responsibility of the world’s creator.

In other respects, however, the limited power and knowledge of human creators may serve to immunize them against common objections against free-will theodicies. In particular, as mentioned earlier, there is little reason to think that the hypothetical creators of virtual worlds will be capable of predicting the future actions of the virtual inhabitants with any great accuracy. So, for example, one common objection to free-will theodicies contends that it should be possible for God to create a world in which each and every human freely chooses to do just those actions that avoid inflicting evil and suffering. If this is true, then the free-will theodicy fails, since the existence of free will does not require evil. Whatever the success of this objection in its original context, however, it fails when applied to virtual worlds. After all, even if such a world is logically possible for an omnipotent God, it clearly falls outside the capacities of human creators. After all, creating such a world would require creators be capable of predicting with precision the future of a virtual world, including all of the choices of all of its virtual inhabitants; moreover, they must do so for an indefinite number of such worlds, in order to find one that marries free will with absence of evil. The capacities of human creators would be presumably be much more limited, amounting to no more than a capacity to predict the general types and magnitudes of evil that virtual agents are likely to suffer in a given virtual world.

Critics of theodicy have also contended that God could miraculously intervene to prevent the sorts of evils that stem from human free will, which would again undermine the free-will theodicy’s claim that evil is an inevitable consequence of allowing humans the capacity to make free choices. Again, however, this objection cannot be easily translated to the case of virtual worlds, at least of the sort we are discussing. So, in the spirit of this objection, let's suppose that we attempt to create a program that continuously monitors the well-being of all of a virtual world's inhabitants, and which intervenes whenever it foresees that a free choice of one inhabitant will cause harm to another inhabitant. We will immediately run into several problems. First, it may be impossible to create such a program, for reasons similar to those pointed out above. After all, it would not be enough for this program simply to predict how the memory states representing the virtual world will evolve over time. In addition, it must correctly identify how these states link up with the emergent properties that we are interesting in tracking, such as the well-being of the various virtual agents. Second, the creation of this sort of world would prevent the user from exercising significant choice, in something like the respect that Swinburne (2004, 240–242) discusses. If the user's reasons for choosing life within the virtual world require that she or he be able to make a difference in the lives of the virtual agents that live there, there is little reason to choose a world with continuous, miraculous interventions that ensure the well-being of these inhabitants.

Given these considerations, what might we conclude about the role of free will in virtual worlds, and what sorts of evil it might serve to justify? First, to the extent that opportunities to exercise free will and autonomy are important for humans, it is reasonable to think this will also hold for human-like virtual agents. This is something that must be accounted for when creating virtual worlds. So, for example, it would seem highly immoral to create a world in which the virtual agents had generally human-like psychologies, and to then create a law of nature that compelled these agents to immediately obey the verbal orders of the user, even in cases where this was in direct contradiction to their most important desires and interests. Second, it is plausible that a virtual world with genuinely autonomous virtual agents will contain some evil, at least in the minimal sense that these agents must be granted some capacity to frustrate the desires of both the user and their fellow inhabitants. Finally, the mere fact that the autonomous agents are virtual does not preclude them from bearing moral responsibility for their actions, or for the possibility that there may some sorts of "deserved" suffering.

Even if all of this is true, however, this fails to show the moral acceptability of creating worlds with the sort of significant evils that some human users of the virtual world might require for their lives to be meaningful. Again, consider the case of a virtual world that resembles, at least in broad respects, a typical combat-heavy video game, where the user must defeat monsters, aliens, or enemy soldiers in order to rescue defenseless children, or members of some other vulnerable group. It might be appropriate, at least within the context of this world, to claim that the villains freely chose to engage in the morally wrong actions they did. However, this does not excuse the creator's moral culpability for making such a world, since she perfectly well could have created a world with more peaceful sociopolitical conditions, or one whose agents were less psychologically prone to violence. Moreover, the free will defense, at least by itself, seems ill-suited to justify the other sorts of evil that might plague a large class of virtual worlds, such as the suffering caused by natural events, or the undeserved harms inflicted on sentient but nonautonomous virtual agents (the digital equivalents of young children and non-human animals). These sorts of concerns also arise in the context of the original problem of evil, of course, but the impossibility of appealing to notions related to libertarian free-will makes addressing them considerably more difficult.

4.2 Natural Laws and Natural Evil

Swinburne (1998; 2004), among others, has appealed to value of simple, uniform natural laws to explain the occurrence of evil that cannot be accounted for by free will in isolation. Swinburne begins with the idea that an important part of leading a worthwhile human life involves the opportunity to make decisions that have a significant impact on the world, for either good or evil. In order to realize goal, he argues that God would need to ensure both that humans' power and knowledge are constrained, and that humans are capable of extending this power and knowledge. He argues that a world (such as the one we live in) with simple, uniform laws of nature that both limit human action and are discoverable by careful inspection of the world meets both criteria. The fact that the invariant, continuous operation of these laws regularly leads to serious instances of both moral and natural evil motivates the agents both to gain knowledge concerning their world, and to take seriously their power to act on it.

Swinburne's claims about the importance of laws of nature for humans might plausibly apply both to the human and virtual inhabitants of virtual worlds. A virtual world without any regular connection between event types, for instance, would make it effectively impossible for its inhabitants to exercise agency in any meaningful way. Similarly, a world in which human users (or others) had unlimited power to continuously reshape the world according to their preferences would undercut the ability of the virtual agents to engage in the sorts of long-term projects that help give substance and structure to their lives. This provides us with a *prima facie* reason for thinking that invariant natural laws ought to be included in a virtual world, even if this entails the existence of evil.

In assessing the ultimate success of this defense when applied to virtual worlds, however, it is important to keep in mind the differences between the way in which an all-powerful God could institute laws of nature, and the ways that humans might design the physics of a virtual world. For instance, on Swinburne's account, God creates a world with simple laws of nature because this is best for the beings inhabiting the world. However, an omnipotent being with different motives could have done otherwise: for example, the laws might appear to be strict regularities whenever intelligent beings were attending to them, but be subject to regular exceptions cases where these would not be noticed, which would allow the creator to pursue various ends undetected. In the case of virtual worlds, by contrast, this does not seem to represent a realistic possibility. For suppose the creators of the world tried to implement an algorithm of this sort, which would allow exceptions to natural laws only in cases where these exceptions would not be detected. This would require that the algorithm consider not only the presence or absence of observers at the time of violation, but also accurately determine whether this violation might contribute to the inhabitants' discovery of the laws' violation over the medium- to long-term. Among other things, this would presumably require representing the state of future science in this world, including the evidence that will be available to these future scientists. The constraints of finite computing power also limit the choice of laws in other ways. For example, the physics underlying the virtual world must be at least tractable, in the sense that a finite computer could use the laws to generate the state of the world at time t_{n+1} from its state at t_n. This need not entail that laws appear simple to the world's inhabitants, of course, since even very simple algorithms may generate highly complex phenomena. However, the constraints of finite computing power rule out certain possibilities that might be open to an omnipotent creator, such as instituting a world with no underlying laws whatsoever, and instead specifying the complete history world as a series of sequential states (already including each of the user's actions), or something of the sort.

Just as was the case with free will, these limitations might actually serve to defuse certain objections to theodicies based on natural laws—e.g., those contending that God might craft exceptions to laws to prevent any and all instances of natural evil. However, the mere fact that a virtual world contains laws need not entail suffering and evil on anything like the scale seen in the real world, even if the laws of nature were somewhat similar. After all, one might prevent large amounts of suffering simply by changing the initial conditions of the virtual world. So, for example, consider the widespread suffering of non-human animals caused by factors such as disease, predation, or starvation, which features prominently in some discussions of the problem of evil[10]. The creators of virtual worlds might significantly reduce or eliminate such suffering by not including certain pathogens or predators, by appropriately limiting rates of reproduction, or by other means. The human creators of virtual worlds also lack many of the resources that theodicies have used to account for natural evil. They cannot, for example, claim that the decisions to incorporate this type of evil were motivated by providing opportunities for what John

[10] For example, Smith (1991) argues that the laws of nature of the real world, and the animal suffering they allow, make it probable that God does not exist.

Hick calls “soul-building” (2010), or for allowing creatures to establish genuine relationships with God. Designing virtual worlds with this capacity is, after all, clearly beyond the scope of limited human creators.

Similar concerns arise when one considers the possibility of providing compensation for those virtual agents whose lives turn out badly through no fault of their own. An all-powerful creator, for example, might arguably be capable of constructing an afterlife that served to fully compensate beings for this sort of undeserved suffering. However, it is implausible that any afterlife within the engineering capacities of human creators could meet this standard. The creators might, for example, provide virtual agents with an afterlife consisting of unending pleasure. However, given the close resemblance between this scenario and Nozick’s original EM, it is highly unlikely that the virtual agents would find this satisfactory. Alternatively, the creators might try to remedy these faults by adopting the strategy described in this chapter, and create yet another virtual world designed to meet the virtual agent’s needs and desires. This, however, leads quickly to an unsustainable proliferation of virtual worlds. So, whatever may be the case with divine creators, human creators cannot compensate virtual agents for the harms done by providing them with an afterlife.

Taken together, these considerations suggest that, while the need to include natural laws might plausibly necessitate the inclusion of some evil within a virtual world, it does not justify the creation of a world with significant amounts of natural or moral evil.

4.3 Hiddenness and Knowledge of the Creator

Many theodicies explicitly address the issue of the “hiddenness” of God, which relates to fact that God’s existence, nature, or both are not readily apparent from an examination of the world around us[11]. Theodicies have attempted to account for this in several ways. First, it might simply reflect the inability of

[11] See McKim (2001, 1–125) for a detailed discussion of the problem.

human knowers to grasp God's transcendent nature. Second, God might choose to remain hidden because human knowledge of God's would limit human freedom, or other important human capacities.

The issue of the hiddenness of the creator presents itself somewhat differently in the context of virtual world, largely because the purposes for which the human creators of virtual worlds might conceal information about themselves are presumably very different from those that theodicies assign to God. Nevertheless, the issue remains a crucially important one. After all, many human users might strongly wish to withhold information about their unique role in a virtual world's creation from its inhabitants, since this knowledge might make it impossible for the user to engage with these other agents as moral equals. Instead, the virtual inhabitants might (with some plausibility) see the human user as god-like figure, to be feared, praised, or blamed, but certainly not to be engaged with as a peer. Similarly, depending on the precise circumstances of the virtual world's creation, it is highly probable that at least some virtual agents would find knowledge of their world's origin to be deeply troublesome, especially since it would seem to trivialize their own attempts to pursue scientific enquiry, construct just social and political institutions, engage in religious practice, and so on.

However, there is little reason to think that these factors by themselves justify such a massive deception of the virtual agents as to the fundamental nature of their world, any more than analogous considerations justify political or religious authorities in suppressing the results of inconvenient scientific investigations. Similarly, while the desire for privacy on the part of the human user may be of some moral importance, it pales in comparison to the importance of allowing virtual agents access to such knowledge. This does not mean, of course, that each virtual agent must be immediately given such knowledge about the virtual world, irrespective of its cognitive or emotional capacities. However, just as parents ought not, in good conscience, systematically deceive their competent, adult children concerning important information about their genetic history, the creators of a virtual world have an obligation to allow the virtual agents means by which they can access the truth about their world.

CONCLUSION

I've argued that our intuitive rejection of life in Nozick's EM is due, at least in part, to our desire to lead a meaningful life. This shortcoming of the EM might be remedied by positing a virtual world that incorporated human-like virtual agents worthy of moral concern, with whom human users might have relationships, and which would enable human users to engage in meaning-giving projects. Moreover, in order to fully satisfy the needs for which typical human users created the worlds, the agents in these worlds would need to suffer significant amounts of evil. Creating worlds of this type, however, raises issues analogous to problem of evil: how could it be morally justifiable to create a world with suffering and evil, when one could have created a better world? I've argued that, while theodicies can be adapted to defend the inclusion of some (minimal) amount of evil, they fail to justify the creation of virtual worlds with significant amounts of evil, such as those modeled on the real world or on contemporary video games.

What might be concluded from all of this? First, to the extent that we think that virtual worlds either currently contain virtual agents worthy of moral concern or that they might eventually come to contain such agents, there are significant moral restraints on the design of such worlds. These worlds must present virtual agents with the opportunity to lead autonomous lives, and to have epistemic access both to the natural laws of their world, and to the world's ultimate nature. These considerations must take place against a background in which the potential suffering of the (not-yet-created) agents is taken seriously, and not simply subjugated to the needs and interests of the human users.

Second, the purposes for which the virtual world is created are relevant to determining the types and magnitude of evil that might be morally appropriate. In this chapter, I've focused on the creation of virtual worlds that might serve as "experience machines" to satisfy certain needs and desires of human users, and have argued that this raises a number of serious moral concerns that cannot be addressed by traditional theodicies, given the divergence between human and divine abilities and motives. It may be, however, that these concerns could be alleviated if the virtual worlds in question were created for some other, more morally significant purpose, such as scientific research that stood to benefit the human or virtual inhabitants of other worlds.

Finally, just as the examination of theodicies can shed light on the tricky problem of what moral concerns might arise in the creation of virtual worlds, it is likely that the creation of such worlds, if it occurs, will provide evidence relevant to assessing the success of problem of evil arguments in the philosophy of religion. The discovery that seemingly satisfactory worlds could be constructed with minimal amounts of evil would, for example, provide some reason for doubting that features such as free will or laws of nature require the sort of evil present in the real world. Conversely, the discovery that it is difficult or impossible for humans to create any virtual world without a significant amount of evil would suggest that claims about the possibility of such worlds—upon which many formulations of the problem of evil are based—may need to be more closely examined.

REFERENCES

Aldiss, Brian. 1969. “Supertoys Last All Summer Long.” *Harper’s Bazaar*.

Anderson, Susan Leigh. 2011. “Asimov’s Laws of Robotics: Implications for Information Technology.” In *Machine Ethics*, edited by Michael Anderson and Susan Leigh Anderson, 285–96. New York: Cambridge University Press.

Belshaw, Christopher. 2014. “What’s Wrong with the Experience Machine?” *European Journal of Philosophy* 22 (4): 573–92.

Benatar, David. 1997. “Why It Is Better Never to Come into Existence.” *American Philosophical Quarterly* 34 (3): 345–55.

———. 2006. *Better Never to Have Been:The Harm of Coming into Existence: The Harm of Coming into Existence*. New York: Clarendon Press.

Bostrom, Nick, and Eliezer Yudkowsky. 2014. “The Ethics of Artificial Intelligence.” *The Cambridge Handbook of Artificial Intelligence*, 316–34.

Crisp, Roger. 2006. “Hedonism Reconsidered.” *Philosophy and Phenomenological Research* 73 (3): 619–45. doi:10.2307/40041013.

Dennett, Daniel C. 1978. “Why You Can’t Make a Computer That Feels Pain.” *Synthese* 38 (3): 415–56.

Dick, Philip K. 1968. *Do Androids Dream of Electric Sheep?* New York: Doubleday.

Floridi, Luciano. 2002. “On the Intrinsic Value of Information Objects and the Infosphere.” *Ethics and Information Technology* 4 (4): 287–304.

Floridi, Luciano, and Jeff W. Sanders. 2004. “On the Morality of Artificial Agents.” *Minds and Machines* 14 (3): 349–79.

Frankfurt, Harry G. 2006. *The Reasons of Love*. Princeton, NJ: Princeton University Press.

Grau, Christopher. 2011. “There Is No ‘I’ in ‘Robot’: Robots and Utilitarianism.” In *Machine Ethics*, edited by Michael Anderson and Susan Leigh Anderson, 461–64. New York: Cambridge University Press.

Gunkel, David J. 2014. “A Vindication of the Rights of Machines.” *Philosophy & Technology* 27 (1): 113–32.

Hewitt, Sharon. 2010. “What Do Our Intuitions about the Experience Machine Really Tell Us about Hedonism?” *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 151 (3): 331–49.

Hick, J. 2010. *Evil and the God of Love*. Palgrave Macmillan UK.

Kawall, Jason. 1999. “The Experience Machine and Mental State Theories of Well-Being.” *The Journal of Value Inquiry* 33 (3): 381–87.

Mackie, J. L. 1955. “Evil and Omnipotence.” *Mind* 64 (254): 200–212.

McKim, Robert. 2001. *Religious Ambiguity and Religious Diversity*. Oxford; New York: Oxford University Press.

Neely, Erica L. 2013. “Machines and the Moral Community.” *Philosophy & Technology* 27 (1): 97–111.

Nozick, Robert. 1974. *Anarchy, State, and Utopia*. New York: Basis Books.

Parfit, Derek. 1976. “On Doing the Best for Our Children.” In *Ethics and Population*, edited by Michael D Bayles. Cambridge, Mass.: Schneckman.

———. 1984. *Reasons and Persons*. OUP Oxford.

Pereboom, Derek. 2005. “The Problem of Evil.” In *The Blackwell Guide to the Philosophy of Religion*, 148–70. Oxford: Blackwell.

Plantinga, Alvin. 1974. *God, Freedom, and Evil*. New York: Harper & Row.

Rachels, Stuart. 1998. “Is It Good to Make Happy People?” *Bioethics* 12 (2): 93–110.

Rowe, William L. 1979. “The Problem of Evil and Some Varieties of Atheism.” *American Philosophical Quarterly* 16 (4): 335–41.

———. 1991. “Ruminations About Evil.” *Philosophical Perspectives* 5: 69–88. doi:10.2307/2214091.

———. 1998. “Reply to Plantinga.” *Noûs* 32 (4): 545–52.

———. 2006. “Friendly Atheism, Skeptical Theism, and the Problem of Evil.” *International Journal for Philosophy of Religion* 59 (2): 79–92.

Salvini, Pericle, Gaetano Ciaravella, Wonpil Yu, Gabriele Ferri, Alessandro Manzi, Barbara Mazzolai, Cecilia Laschi, Sang-Rok Oh, and Paolo Dario. 2010. “How Safe Are Service Robots in Urban Environments? Bullying a Robot.” In *RO-MAN, 2010 IEEE*, 1–7. IEEE.

Smith, Quentin. 1991. “An Atheological Argument from Evil Natural Laws.” *International Journal for Philosophy of Religion* 29 (3): 159–74.

Sober, Elliot. 2000. “Psychological Egoism.” In *The Blackwell Guide to Ethical Theory*, edited by Hugh LaFollette, 129–48. Malden, MA: Blackwell.

Sumner, L. W. 1996. *Welfare, Happiness, and Ethics*. New York: Clarendon Press.

Swinburne, Richard. 1998. *Providence and the Problem of Evil*. New York: Oxford University Press.

———. 2004. *The Existence of God*. 2nd ed. Clarendon Press.

Tooley, Michael. 2015. “The Problem of Evil.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2015. http://plato.stanford.edu/archives/fall2015/entries/evil/.

Torrance, Steve. 2007. “Ethics and Consciousness in Artificial Agents.” *AI & SOCIETY* 22 (4): 495–521. doi:10.1007/s00146-007-0091-8.

Van Inwagen, Peter. 2008. *The Problem of Evil*. New York: Oxford University Press.

Williams, Bernard. 1973. “The Makropulos Case: Reflections on the Tedium of Immortality.” In *Problems of the Self: Philosophical Papers 1956-1972*, 82–100. Cambridge, UK: Cambridge University Press.

Wolf, Susan, Stephen Macedo, John Koethe, Robert M. Adams, Nomy Arpaly, and Jonathan Haidt. 2012. *Meaning in Life and Why It Matters*. Princeton, N.J.: Princeton University Press.