Journal of Virtual Worlds Research

jvwr.net

ISSN: 1941-8477

The Dark Side of Virtual Worlds

March 2020

Volume 13 No. 1

Cover: pexels-photo-1851243-frank

Volume 13, Number 1

The Dark Side of Virtual Worlds

March 2020

Cover image: pexels-photo-1851243-frank

Journal of Virtual Worlds Research jvwr.net
ISSN: 1941-8477

Volume 13, Number 1
The Dark Side of Virtual Worlds
March 2020

Artificial Beings Worthy of Moral Consideration in Virtual Environments: An Analysis of Ethical Viability

Stefano Gualeni
Institute of Games – University of Malta, Malta

Abstract

This article explores whether and under which circumstances it is ethically viable to include artificial beings worthy of moral consideration in virtual environments. In particular, the article focuses on virtual environments such as those in digital games and training simulations – interactive and persistent digital artifacts designed to fulfill specific purposes, such as entertainment, education, training, or persuasion.

The article introduces the criteria for moral consideration that serve as a framework for this analysis. Adopting this framework, the article tackles the question of whether including artificial intelligences that are entitled to moral consideration in virtual environments constitutes an immoral action on the part of human creators. To address this problem, the article draws on three conceptual lenses from the philosophical branch of ethics: the problem of parenthood and procreation, the question concerning the moral status of animals, and the classical problem of evil.

Using a thought experiment, the concluding section proposes a contractualist answer to the question posed in this article. The same section also emphasizes the potential need to reframe our understanding of the design of virtual environments and their future stakeholders.

1. Introduction

Who or what is entitled to moral consideration? The criteria for inclusion in a 'moral community' typically involve a broad and interconnected array of socio-cultural factors and, as such, the question of ethical relevance cannot be expected to be answered in a univocal or exhaustive manner. In the history of Western thought, criteria for ethical relevance have often relied on the detectable presence of certain cognitive and intellectual capabilities[1]. Criteria based on the ability to exhibit awareness and rationality, however, exclude specific groups of human beings from being considered morally relevant. Infants, severely mentally disabled people, individuals with Alzheimer's disease, and people in comas are often mentioned as problematic cases begging the question of how rational a rational being must be to be recognized as worthy of moral consideration (see Bostrom Bostrom & Yudkowsky 2014, p. 6; Neely 2014, p. 2).

Dissatisfied with inherently anthropocentric criteria based on cognitive and intellectual capabilities, thinkers such as Taylor (1996) and Singer (2002) have recommended using the principle of *sentience* to determine whether a being should be given moral consideration. Using sentience to establish moral inclusion would ultimately mean arguing that if something is capable of feeling pain or discomfort, it would be ethically wrong to cause that thing any amount of unnecessary suffering[2].

With the objective of extending the applicability of such a sentience-based approach, this article adopts the idea that the necessary condition for moral relevance is not sentience per se, but rather having 'interests.' According to Basl's (2012, pp. 4–5) definition, for example, a being's interests "are those things the satisfaction of which contributes to its well-being." In her work on ethical analysis, Neely has proposed focusing on the specific interests expressed by an entity concerning preserving its own autonomy and bodily integrity, where the latter concept refers to the possibility of continuing one's existence undisturbed and unharmed. From this standpoint, she argues, suffering is a way of expressing (and detecting) those interests that are specific to biological beings (Neely 2014, p. 3).

At this juncture, it is important to point out that Neely does not precisely define what she means by 'autonomy.' In line with common use and the Greek etymology of the term, one can assume that she is referring to the capability for self-regulation – the condition of being free from external control or influence. As the concept of autonomy is central to the present work, it is desirable to frame the term in a way that is less implicit and nominal. I thus propose defining 'autonomy' not in terms of the absence of or liberation from limitations and interdictions, but as the possibility for an individual to determine and fashion himself or herself in relation to the limitations and interdictions that characterize being in a world (regardless of its actual or virtual constitution).

Following Neely, a being expressing interest in maintaining its autonomy and integrity will be used in this text as the basic criterion for worthiness for moral consideration. This decision was motivated by the greater inclusivity of this approach compared with other ways of assessing moral relevance. A second reason for adopting this interest-based criterion for moral inclusion is that, because this approach moves beyond a bio-centric conception of ethical relevance, it can be directly

[1] As a case in point, Socratic ethics has frequently been characterized by scholars as 'intellectualist' because it treated questions and issues concerning ethics and virtue as matters of knowledge and self-knowledge. In Plato's writings, Socrates argued that only our rational faculties – not emotions or instincts – should determine our ethical decisions (Frede 2017). In the *Nicomachean Ethics*, Aristotle similarly claimed that, among living things, only humans are responsive to reason and that it is precisely because of these intellectual capabilities that humans can bypass or suppress emotions and pursue virtues (Aristotle 1999).

[2] Here, 'unnecessary' is to be understood in relation to Singer's (2002) utilitarian position that there are situations in which the pursuit of the greater utility for the moral community can make the deliberate use of violence permissible.

applied in the ethical consideration of artificial intelligences[3]. From this standpoint, the concept of ‘suffering,’ which conventionally guides sentience-based approaches to ethics, can be replaced by the more inclusive notion of ‘damage’ (see Textbox).

In the case of artificial persons (that is, general artificial intelligences that can be considered worthy of ethical consideration), the very notion of a ‘body’ employed by Neely in her articulation of what constitutes interests becomes ambiguous and potentially exclusionary. Her approach could be made more inclusive and useful for the goals of the present inquiry by positing that worthiness for ethical consideration, requires, at minimum, that a being somehow expresses an interest in continuing to exist, regardless of whether or not they have a physical body.

‘Damage’ does not simply refer to physical harm or the impairment of certain functionalities. Rather, in this discussion, ‘damage’ indicates the disrespect of the bodily integrity and/or the independence of an autonomous moral agent. Certainly, an electric wheelchair or the battered laptop I am using to write this article cannot be considered either autonomous or moral because they lack psychological interests as well as the ability to formulate preferences and goals for themselves (see Basl 2012).

2. On Evil and the Moral Consideration of Artificial Beings:

Evil is generally defined as the most comprehensive expression of moral disapproval (Floridi & Sanders 2001, p. 55). Philosophical reasoning on ethics has traditionally distinguished two types of evil, namely ‘moral evil’ (ME) and ‘natural evil’ (NE).

- ME presupposes moral agents who, having sufficient information about a given situation, act autonomously and intentionally. In other words, ME represents the kind of disapproval that one can direct towards agents who can be considered morally responsible for their unethical actions.
- NE, in contrast, refers to the actions of non-autonomous agents, as in the case of natural disasters such as earthquakes or floods, for example. NE thus often indicates a kind of evil that arises independently of human intervention.

Floridi and Sanders (2001, 59) have observed that advancements in the development of science and technology are progressively blurring the boundaries between ME and NE:

> [I]n advanced societies, people are confronted by visible and salient evils that are neither simply natural nor immediately moral: an innocent dies because the ambulance was delayed by the traffic; a computer-based monitor ‘reboots’ in the middle of surgery because its software is not fully compatible with other programs also in use, with the result that the patient is at increased risk during the reboot period.

Evil actions that involve both moral agents (e.g., ambulance drivers, surgeons) and non-morally accountable agents (e.g., the traffic, the computer-based monitor) are difficult to categorize unambiguously in terms of the classical ME/NE separation. After all, the classical conceptualization of evil was developed in an era when technology was understood as a mere instrument for human practices, and moral concerns were framed solely in terms of human or divine responsibility. As the

[3] In this article, the phrase ‘artificial intelligences’ refers to beings whose intelligence did not emerge spontaneously (i.e. without being designed and manufactured by human beings) or develop on an organic substrate. To be recognized as ‘intelligent’, artificial agents need to be generally intelligent, in the sense that their intelligence must not be task-specific (see Bostrom & & Yudkowsky 2014, pp. 3–6; Togelius & Yannakakis 2016).

above indeterminate situations described by Floridi and Sanders demonstrate, such conceptualization has become obsolete.

Notably, Floridi and Sanders (2001) implicitly limit their claims to evil actions that are actual (i.e., occurring in the actual world). More specifically, these authors conceptualize and exemplify moral agency in a way that refers to actions performed in the actual world (i.e., the world we share as biological organisms), meaning that their understanding of moral agency does not involve actions that are carried out within the virtual worlds of digital games, for example. Consequently, in their 2001 article, Floridi and Sanders do not address the specific kinds of behaviors and evils for which artificial intelligences in virtual worlds could be considered moral agents or understood as beings that are worthy of moral consideration. In a later essay ('On the morality of artificial agents'), Floridi and Sanders (2004) do discuss the possibility of artificial intelligences being entitled to ethical consideration; however, this discussion is cursory and instrumental to framing the moral accountability of human beings in their roles as users and designers of software.

The exclusive focus on the ethical stance of human agents in their relationship with digital media is symptomatic of what Bostrom and Yudkowsky (2014, p. 7) refer to as 'the wide agreement that current AI systems have no moral status.' As an example, Bostrom and Yudkowsky point out that, in our current stage of technical development, we customarily modify, copy, terminate, delete, and use artificial intelligences or parts of them without considering these actions as having moral implications.

The reflections presented in this article attempt to answer the question of whether and under which circumstances the use of artificial intelligences worthy of moral consideration in interactive virtual environments constitutes an evil act on the part of human beings. 'Virtual environments' can be understood as a particular kind of digital artifact (e.g., the operating theatre of a surgical simulation or a level in a digital game) that can be experientially engaged by any kind of agent as a world, that is to say both phenomenologically and existentially (see Vella & Gualeni 2019).

3. On Actual Evil in Virtual Environments

In this section, I illustrate my argument through a thought experiment. In this speculative exercise, I ask the reader to imagine a situation in which artificial moral agents worthy of moral consideration are not only technologically accomplished, but already actively inhabit the virtual environments of digital games, training simulations, interactive archaeological reconstructions, and so forth. Assuming that general artificial intelligences entitled to moral consideration are technologically possible (see Footnote 3), is it morally viable to use these artificial beings in virtual worlds? In other words, is it an evil act for human beings to use artificial intelligences worthy of moral consideration in virtual environments designed exclusively to serve human goals (e.g., entertainment, education, persuasion, and training)?

As they are currently designed, the virtual environments of digital games and training simulations are characterized by the risk of causing damage to their artificial inhabitants. In the Western world, for instance, we use interactive digital models to familiarize ourselves with situations and procedures that are hazardous in the actual world. It is well established in the literature that, through the iterative manipulation of virtual scenarios, users can develop several forms of knowledge that can be partially or fully transferred to actual situations without damaging actual equipment or putting the welfare or the existence of other humans at risk (see Vella & Gualeni 2019, p. 125; Silcox 2019, p. 75). This refers, for instance, to digital simulations that allow users to experience the interactive performance of particularly difficult landing procedures for commercial airliners, or high-risk counter-terrorist operations, for example. In these virtual environments, artificial intelligences are, by design, limited in their agency and stunted in their autonomy for the sake of the intended experiential and didactic goals of the interactions.

At face value, it is clearly not morally acceptable to use violence against artificial intelligences that can be considered worthy of moral consideration; deliberately damaging or hurting such beings is, by definition, an immoral act. The examples outlined above can be considered extreme cases in which humans, as both creators and users, are already exerting violence directly and explicitly on artificial intelligences. Less obvious scenarios, where highly creative and competitive artificial moral agents are used in a non-violent digital gameworld intended for human enjoyment, can also be considered. Possible examples could include a digital game with rules analogous to tennis or pool. In the persistent gameworld of such digital games, artificial moral agents are forced to measure their skills against human players. Is it morally permissible, in that context, to limit how those general artificial intelligences can learn and perform for the purpose of making the experience of the game enjoyable for human players? Is it an act of evil to stunt the artificial agents' autonomy and aspirations to ensure that their skill levels remain comparable to those of the human players?

Shea (2017, p. 144) has hypothesized that artificial intelligences experiencing scenarios like those outlined above 'might find their lives to be worth living, at least in some minimal sense. Nevertheless, they might find the situation deeply dispiriting and frustrating'. On the basis of his observations, and as part of the overarching argument presented here, it is meaningful to question whether a developer who creates virtual environments that are structurally unfair to artificial moral agents can be considered morally responsible. Questions concerning the ethical viability of creating virtual environments that are systematically oppressive or unfair to some of their inhabitants have several analogies with existing scholarly themes and positions, some of which were examined by Shea (2017). Likewise, the question of whether it is moral to use artificial beings worthy of moral consideration in these environments can also be linked to existing conceptual perspectives. Three of these perspectives are particularly relevant to the main question posed in the present article:

- The theme of animal oppression and the moral status of non-human animals (Section 3.1);
- The ethical questions concerning parenthood and procreation (Section 3.2); and
- The classical problem of evil (Section 3.3).

3.1. Animal Oppression and the Moral Status of Non-Human Animals

The subordination of non-human animals to the needs and desires of human beings is often referred to as 'speciesism,' a term coined by Ryder in the 1970s to denote an anthropocentric prejudice analogous to racism (Ryder 2000). In its most extreme form, speciesism excludes any species other than *Homo sapiens* from taking part in our moral circle. The actions and attitudes of speciesists are deemed biased because they lack moral justification for the preference and prioritization of the interests of human beings over those of other species (Gruen 2017).

For those who oppose speciesism, any action that fails to treat an animal as inherently worthy of moral consideration is ethically objectionable. According to the animal rights movement, for example, treating an animal as a means to some human end is immoral; using animals in circuses and experimenting on them in medical research are thus considered obvious examples of evil. The utilitarian position on animal rights, in contrast, argues that an animal's moral significance depends on the competing moral claims that are relevant in a given situation (Gruen 2017). This position, which is commonly associated with Singer (2002), asserts that, although the interests of all beings worthy of moral consideration are of equal importance, it is not necessarily morally wrong to violate or frustrate some of those interests. As such, a utilitarian position would not condemn *a priori* the possibility of restraining or exploiting artificial intelligences that are worthy of moral consideration, as the damage caused to these beings might be an acceptable price to pay for the achievement of wider – and more pressing – goals of the moral community.

Without attempting to provide a complete overview of the various questions and positions involved in the problem of the moral status of animals, I would like to highlight several conceptual similarities in the ethical problems involved in animal oppression and in the subjugation of artificial intelligences to human needs and preferences. Discussions of the moral status and autonomy of non-human animals resonate with the conversations currently taking place in the context of the ethics of artificial intelligence. In recent years, a growing body of theoretical literature has emerged on the personhood of non-human animals, especially non-human great apes, dolphins, and elephants, and the legal rights that should be accorded to them (see e.g., Cavalieri & Singer 1996; DeGrazia 1997; Ross 2019). Similarly, the nascent field of the ethics of artificial intelligence advocates for 'robot rights' and the fair treatment of artificial intelligences, leveraging a less anthropocentric understanding of moral significance (see e.g., Coeckelbergh 2010; Gunkel 2018).

Although several countries such as New Zealand, Germany, and Spain now grant certain rights to some non-human animals, to date, no legal steps have been taken to protect artificial beings that could be recognized as worthy of moral consideration. The ethics of artificial intelligence is a relative newcomer in discussions concerning morality beyond the immediate interests of human beings. Additionally, because of existing artificial beings' lack of sophistication, the general population may not yet deem it pressing or particularly intuitive to consider these beings worthy of ethical consideration. Nevertheless, many scholars and researchers consider it worthwhile to speculate on and prepare for the expansion of our moral circle in the direction of artificial beings. This drive is evident in recent publications proposing frameworks for personhood and ethical consideration that can include humans, non-human animals, embryos, and artificial intelligences (see e.g., Neely 2014; Kurki & Pietrzykowski 2017).

In sum, if we do not ascribe to an anthropocentric and utilitarian view on the matter, holding animals or general artificial intelligences captive and/or violating their autonomy by forcing them to take part in heteronomous activities is to be considered morally wrong.

3.2. Parenthood and Procreation

Is it ethical to knowingly generate children who will be less well-off than children resulting from other decisions? Is it morally wrong for people to choose to reproduce when they have reason to believe that their children will live in a way that the potential parents deem sub-optimal? These questions concerning parenthood and procreation rely both on the parents being sufficiently well-informed about certain states of affairs and on their understanding of what constitutes a 'good life.'

The biological process that generates a child can be recognized as, to some degree, analogous to the process of technologically developing a general artificial intelligence that is worthy of moral consideration (see the 'principle of ontogeny non-discrimination' in Bostrom 2011, pp. 6-7). Building on that premise, I argue that the decision to bring a child into the actual world and the decision to create an artificial being worthy of moral consideration in a virtual environment are both non-trivial ethical choices. From a moral standpoint, the creation of such a general artificial intelligence might even be more ethically problematic than becoming a parent: Software developers have a higher degree of control over their creations than parents have in human biological reproduction. As digital creators, software developers can make decisions concerning the production of both virtual environments and the artificial autonomous beings inhabiting these environments, whereas parents can, at best, play an active role in the production of the child.

Inspired by the framing of decisions regarding procreation in the context of classical ethics, this section focuses on whether the information one has about the well-being of hypothetical artificial beings that are worthy of moral consideration could ever make it ethically viable to use these beings in a persistent virtual environment designed to meet specific human needs and desires.

In the context of the above example, it is reasonable to suppose that the human developers responsible for the creation of a tennis or pool digital gameworld have extensive knowledge about their creations. At a minimum, these developers are aware of whether using artificial intelligences in these specific persistent digital gameworlds will expose these artificial moral agents to direct acts of violence or to the risk of having their existence forcefully discontinued. By the same token, the developers are also cognizant that their creations will be bound to a virtual existence characterized by a large degree of dissatisfaction. Using these artificial intelligences in digital gameworlds and forcing them to repeat behavioral patterns can be recognized as a form of slavery that is not unlike the captivity suffered by, for example, circus animals (cfr. Section 3.1).

Shea (2017) described what is perhaps an even more problematic case of deliberate limitation of an artificial moral agent's autonomy. He observed that sections of a virtual environment are only activated when a user decides to visit them. Supposing this is the case for the virtual environments we are considering, then the artificial intelligences needed in a certain scenario or situation might also be forcefully activated, deactivated, or removed in response to the decisions of a human user (Shea 2017, p. 143).

Clearly, being held captive, being coerced into frustrating subservience, and having one's existence forcefully discontinued are not elements of anyone's definition of a 'good life.' On the contrary, some degree of autonomy is among the necessary preconditions for all classic philosophical understanding of a 'good life.' In a way that resonates with arguments against slavery and animal oppression, it must be concluded that it is not morally viable to use autonomous artificial intelligences in environments where, by design, their autonomy will largely be suppressed.

Although this conclusion appears obvious, it must be acknowledged that there are several exceptions to this point. For example, following a Millsian approach[4], one could argue that it would not be immoral to limit the autonomy of agents who behave in ways that may cause damage to themselves or to other morally relevant beings (e.g., imprisoning a convicted felon or putting a dog on a leash in public). Another potential exception to the moral impermissibility of limiting the autonomy of artificial moral agents is reflected in the following question: Would it be morally wrong to act against the welfare of artificial moral agents or to systematically limit their autonomy if these artificial intelligences were programmed not to feel dissatisfied with this practice, or even to masochistically enjoy this kind of heterotelic imposition? (See also Neely 2014, p. 2 and Shea 2017, p. 143). I argue that this question is only problematic at first glance, because the limitation of autonomy outlined above is more comparable to a lobotomy than to, for instance, a vaccination. In other words, instead of being a way to preserve an agent's long-term well-being and fulfill their interest in self-preservation, such a limitation of autonomy would be an imposition of constraints on the artificial intelligences' potential to fully express their autotelic interests and to experience a vast spectrum of existential situations. I consider this limitation incompatible with what could be considered 'moral design' from the developer's point of view[5].

[4] Here, we are referring here to John Stuart Mill's understanding of paternalism within a liberal framework for ethics. For more information, see his 1859 essay *On Liberty* (Mill, 1991).

[5] At this point, it may be useful to recall Aldous Huxley's 1932 novel *Brave New World*, in which one of the central themes is the question of whether it is preferable to be happy or autonomous. Specifically, we could ask under which conditions it would be morally viable to produce citizens in the 'epsilon class' found in Huxley's fictional world. This class of human beings was deprived of oxygen during their artificial gestation, keeping their mental capabilities to a minimum functioning level so that they would not be frustrated by the monotonous labour required of them as adults.

3.3. The Classical Problem of Evil

The kinds of behaviors that are expected and encouraged in most contemporary virtual environments raise significant ethical concerns. The ubiquity of violence and the lack of nuance in the ethical choices afforded to the player are only some of the most evident moral problems characterizing contemporary digital game production. Additionally, as discussed specifically in Section 3.2, even supposing that the user's behavior in a certain virtual environment is completely moral, there are still ethical dilemmas inherent in the inclusion of artificial intelligences worthy of moral consideration in virtual environments. This is because all virtual environments created to serve human purposes will inevitably lead to significant damage and frustration for the artificial moral agents that inhabit them. From the hypothetical perspective of these artificial moral agents and in analogy with the classical 'problem of evil', whoever created the environment that these beings inhabit can be considered morally responsible for the systemic and widespread suffering that they endure.

Several analogies can be drawn between the question motivating this article and religious approaches to the problem of evil. Clearly, there are many important differences between the roles and possibilities of human creators and those of a divine demiurge. First, despite having the capability to create artificial, experiential worlds and maintaining considerable control over their digital media creations, humans are neither omnipotent nor omniscient. Their knowledge about the technologies they use is always incomplete and in progress, as are the ways in which individuals and societies use and attribute meaning to technological creations. This means that, even when human developers are motivated by benevolent intentions in creating virtual environments, their lack of foresight and their inherently limited technical capabilities make it impossible for them to envisage virtual experiences that are completely free of potential sources of damage or even environments that systemically optimize the well-being of the moral agents that inhabit them.

With these differences in mind, a version of the classical problem of evil that is specific to human beings as creators of virtual environments can be formulated.

a. All virtual environments created for human purposes will lead to significant damage and frustration for the artificial moral agents worthy of moral consideration that inhabit them.

b. It is morally wrong to create virtual environments that potentially cause damage and frustration when one has the power and the knowledge to create environments that reduce these effects.

c. Developers are capable of reducing the damage and frustration caused by the virtual environments they create, for example, by creating environments that do not include artificial moral agents worthy of moral consideration.

d. It is morally wrong to use artificial moral agents in virtual environments that would cause these agents damage and frustration.

We are once again confronted with the idea that, if we are willing to adopt a sufficiently wide conceptualization of a moral community, including artificial intelligences in virtual environments that are exclusively designed for human uses and goals would not be morally permissible.

4. Conclusions

From the perspectives offered by the three conceptual lenses on ethics discussed in Section 3, there does not seem to be a way for humans to design and develop virtual environments that are not – or even that are not systematically – evil *vis-à-vis* the artificial moral agents that inhabit them. In this final section, I propose several potential characteristics for virtual environments that would enable them to meet the needs of human users and respect the autonomy of artificial beings that are recognized as worthy of moral consideration.

In the introduction to this article, ‘autonomy’ was defined as the possibility for an individual to shape himself or herself in relation to the limitations and interdictions that characterize being in a world, an understanding tied to the existential conception of freedom articulated by Sartre (1966) in *Being and Nothingness*. This understanding of autonomy serves to defend divine providence in view of the existence of evil, attributing the responsibility for evil actions to moral agents themselves rather than to their heavenly creator. This definition of autonomy also accounts for the (artificial or biological) agent’s potential to decide to take part in or to withdraw from a world; Camus (1955, p. 3) framed this fundamental aspect of existential freedom as follows in the well-known opening of his *The Myth of Sisyphus*: ‘[t]here is only one really serious philosophical question, and that is suicide’.

The framing of autonomy adopted in this article is crucial for its final point. Until now, I have discussed the moral impermissibility of including artificial beings worthy of moral consideration in virtual environments exclusively from the point of view of the human developers. Would this use of artificial intelligences still be ethically impermissible if they were given the possibility to autonomously decide whether to take part in specific virtual environments and when to abandon them? In this scenario, artificial intelligences could be presented with epistemic access to the information and characteristics of each virtual environment in which they could take part, as well as details concerning the activities expected to take place within these environments. Thus, the artificial intelligences would not be heteronomously forced into a condition of slavery; instead, they would be allowed to act in virtual environments with levels of knowledge and autonomy comparable to those of human users. In this hypothetical situation, autonomous artificial intelligences would relate to virtual environments not only as moral agents but effectively as existential subjects (see Vella & Gualeni 2019). Following this speculative trajectory, the only possible positive answer to the question regarding the ethical viability of using artificial intelligences worthy of moral consideration in virtual worlds relies upon granting these artificial moral agents a wider, existential kind of autonomy in virtual environments.

As a possible answer to the question of under which circumstances it would it be morally viable to include beings worthy of moral consideration in a virtual environment, I propose a variation of ‘contractualism,’ a form of moral reasoning based on a contract or agreement among the autonomous moral agents involved in a certain situation or activity[6]. This solution could perhaps be seen as downplaying the ethical responsibility of the creators of digital games and training simulations. This observation notwithstanding, I believe that, through their design decisions in the virtual environments they create, developers can – and likely should – be held accountable for enabling, facilitating, or directly causing the beings inhabiting these virtual environments to experience damage and frustration. In classical terms, these developers would be deemed responsible for all NEs in the virtual environments they create. Ethical practices, however, are not simply or exclusively a matter of direct agency: Actions and behaviors are always invited, mediated, discouraged, delegated, and modified by our technological devices and systems (see Verbeek 2011; Gualeni 2015). In view of these observations, we should consider the designers’ responsibilities and how their ethical and ideological positions translate (consciously or less so) into material affordances and teleologies in discussions of various types of moral scenarios.

[6] It is relevant to emphasize that a contractualist stance is fundamentally dependent on the notion of ‘moral personhood’ and that this perspective recognizes an equal moral ground for all the moral persons participating in the agreement. Unlike the utilitarian position briefly outlined in this text, however, contractualism does not claim that there is only one rational attitude regarding value.

References

Aristotle (1999). *Nicomachean ethics* (W. D. Ross Trans.). Kitchener, ON: Batoche Books (Original work published ca. 350 B.C.E.).

Basl, J. (2012). Machines as moral patients we shouldn't care about (yet): The interests and welfare of current machines. In Gunkel, D. J., Bryson, J. J., & Torrance (eds.). *Proceedings of the AISB/IACAP World Congress 2012: The Machine Question: AI, Ethics and Moral Responsibility*. Birmingham, UK.

Bostrom, N. & Yudkowsky, E. (2014). The ethics of artificial intelligence. *The Cambridge handbook of artificial intelligence*, *1*, 316-334.

Camus, A. (1955). *The myth of Sisyphus and other essays*. New York, NY: Alfred A. Knopf.

Cavalieri, P. & Singer, P. (Eds.) (1996). *The great ape project: Equality beyond humanity*. New York, NY: Macmillan Publishers.

Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and information technology*, *12*(3), 209-221.

DeGrazia, D. (1997). Great apes, dolphins, and the concept of personhood. *The Southern Journal of Philosophy*, *35*(3), 301-320.

Floridi, L. & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, *14*(3), 349-379.

Floridi, L. & Sanders, J. W. (2001). Artificial evil and the foundation of computer ethics. *Ethics and Information Technology*, *3* (1), 55-66.

Frede, D. (2017, 6 December). Plato's Ethics: An overview. *The Stanford Encyclopedia of Philosophy Archive* (Winter 2017 Edition). Retrieved from https://plato.stanford.edu/archives/win2017/entries/plato-ethics

Gruen, L. (2017, 23 August). The moral status of animals. *The Stanford Encyclopedia of Philosophy* (Fall 2017 Edition). Retrieved https://plato.stanford.edu/archives/fall2017/entries/moral-animal/

Gualeni, S. (2015). *Virtual worlds as philosophical tools*. Basingstoke, UK: Palgrave Macmillan.

Gunkel, D. J. (2018). *Robot rights*. Cambridge, MA: The MIT Press.

Huxley, A. (2008). *Brave new world*. Stuttgart, Germany: Ernst Klett Sprachen.

Kurki, V. A., & Pietrzykowski, T. (eds.) (2017). *Legal personhood: Animals, artificial intelligence and the unborn*. New York, NY: Springer.

Mill, J. S. (1991). On liberty. In J. Gray, (ed.), *John Stuart Mill: On Liberty and Other Essays*. Oxford, UK: Oxford University Press. (Original work published 1859).

Neely, E. L. (2014). Machines and the moral community. *Philosophy & Technology*, *27*(1), 97-111.

Ross, D. (2019). Consciousness, language, and the possibility of non-human personhood: reflections on elephants. *Journal of Consciousness Studies*, *26*(3-4), 227-251.

Ryder, R. D. (2000). *Animal revolution: Changing attitudes towards speciesism*. Oxford, UK: Berg Publishers.

Sartre, J. P. (1966). *Being and nothingness*. (H. E. Barnes Trans.). New York, NY: Washington Square Press.

Shea, B. (2017). The problem of evil in virtual worlds. In M. Silcox (Ed.) *Experience machines: The philosophy of virtual worlds* (pp. 137-154). London, UK: Rowman and Littlefield International.

Silcox, M. (2019). *A defense of simulated experience: New noble lies*. New York, NY: Routledge.

Singer, P. (2002). *Animal liberation* (3rd ed.). Ecco, USA.

Taylor, A. (1996). Nasty, brutish, and short: The illiberal intuition that animals don't count. *The Journal of Value Inquiry, 30*, 265-277.

Togelius, J., & Yannakakis, G. N. (2016). General general game AI. *2016 IEEE Conference on Computational Intelligence and Games (CIG),* (pp. 1-8). IEEE.

Vella, D. & Gualeni, S. (2019).Virtual subjectivity: Existence and projectuality in virtual worlds. *Techne': Research in Philosophy of Technology, 23*(2).

Verbeek, P. P. (2011). Moralizing technology: Understanding and designing the morality of things. Chicago, IL: The University of Chicago Press.