Journal of EVOLUTION & Technology

A peer-reviewed electronic journal published by the Institute for Ethics and Emerging Technologies

ISSN 1541-0099

29(2) – November 2019

Machines and Non-Identity Problems

Zachary Biondi
Department of Philosophy, UCLA

zbiondi@humnet.ucla.edu

Journal of Evolution and Technology - Vol. 29 Issue 2 – October 2019 – pgs 12–25

Abstract

A number of thinkers have been wondering about the moral obligations humans have, or will have, to intelligent technologies. An underlying assumption is that "moral machines" are decades in the offing, and thus we have no pressing obligations now. But, in the context of technology, we are yet to consider that we might owe moral consideration to something that is not a member of the moral community but eventually will be as an outcome of human action. Do we have current actual obligations to technologies that do not currently exist? If there are obligations to currently non-existing technologies, we must confront what might be called the Non-Identical Machines Problem. Can we harm or benefit an entity by making it one way rather than another? This paper presents the problem and argues that it is more challenging than the standard Non-Identity Problem.

Introduction

For philosophers, issues in artificial intelligence (AI) have historically been situated in the philosophy of mind. A renewed interest is less in the traditional question of whether artificial general intelligence is possible and more in the ethical problems presented by practically capable technologies, whether or not they are conscious. The concern falls out of the fact that we are hurtling with ever increasing speed toward more and more powerful types of technology.

A number of scientists and scholars have also been wondering about the moral obligations that humans might have to emergent technologies. If we develop an artificial general intelligence, would it be possible to harm it? We usually assume that items of technology that belong in the moral community (what we might call *moral machines*) are decades in the offing, and thus we have no obligations toward them now (Wallach and Allen 2009). One day, however, when they reach a certain level of sophistication, the obligations will take effect.

I worry that the discussion is ignoring a different obligation. If so, we risk a different moral failure. I am interested in what I call the problem of *creative responsibility*: what responsibilities does a creator have to its creation *in the act of creating*? To help draw out the contours of the problem, I will relate it to the two more dominant themes within the ethics of future technology that are mentioned above: 1) existential risk/safety; and 2) the moral machine problem (Biondi 2018b).

The second of these topics does not involve moral issues that precede the created entity's existence. It is about how we should categorize and behave toward entities we find among us. The first, however, is about moral issues that we face prior to the existence of the technologies in question. Yet the issues are about the technologies only indirectly. The moral concern is the (types of) entities we find among us.

We thus find conceptual space for a third topic: moral issues that we face prior to the existence of the created entity and in which the focus of moral concern is the created entity itself. The problem of creative responsibility is about how we confront these issues. (There is also the meta(ethical) problem – which I will not directly address here – of how to make the problem solvable, or even intelligible, within canonical ethical theories.)

To frame my discussion further, there are two ways to construe the problem of creative responsibility:

1. ***Narrow***. We can wonder about our responsibilities in cases in which we are creating an entity that is owed moral concern. Here we are assuming that the entity in question is a member of the moral community, or will be when it exists. There is a further distinction to draw (note that a single case can have aspects of both):

    a. ***"When" Choices***. These are situations in which the choices are about whether, when, or in what environment to create an entity that is not itself the subject of creative design choices. These problems have familiar literatures. Examples would be the morality of having children (Benatar 2008) or breeding nonhuman animals.

    b. ***"How" Choices***. These are cases in which the moral entity is the subject of creative design choices. How or in what manner the entity will exist (its *nature*, so to speak) is to be determined by the creator. The problems here have been discussed with respect to genetic/neuro/moral enhancement (Savulescu 2009), disability, and domestication. To illustrate, Yahweh in the Book of Genesis faced "how" choices of creative responsibility on the fifth and sixth days of creation.

2. ***Broad***. We can wonder about the ethics of creation generally: for example, the choice to create or not create *at all*, choices about the creation of environments, and the choice of whether to create an entity with moral significance or one without. How should we conceive of our eventual creations, knowing that the conceptions will guide our creative acts? Yahweh faced this problem in the beginning, and each day after. The choice to build an AI in the first place is a matter of broad creative responsibility, and thus we must not beg any questions by assuming that *not* creating an AI is morally safe (though it likely is from the perspective of narrow creative responsibility). The question of which machine we should make is, or can be heard as, the *broad* problem since many of the possible types of entities, we assume, would not be entities owed moral concern. (See Plato's *Timaeus* 29e-30c for a statement of the broad construal.)

As best as I can, I will set the broad construal aside. However, assumptions about broad creative responsibility necessarily lurk in the background of my discussion (and any philosophical project *qua creation*, in fact). My plan is to focus on issues of *narrow* creative responsibility. The narrow construal

connects to what Derek Parfit called the Non-Identity Problem (NIP) (see Parfit 1984, ch. 16, and Parfit 1976; Kavka 1982; Adams 1979; Schwartz 1978; Adams also circles around the NIP in an earlier contribution (1972)). The "when" choices run into the standard NIP: your choices can't make a particular future individual better or worse off because such choices change the identity of the individual in question. "How" choices yield more challenging versions of the NIP. With my focus on technology, we might coin the term *Non-Identical Machines Problem* (NIMP): since design choices will affect which type of machine exists, the choices are unable to make the machine better or worse off. And, as an important complicating factor, we have already begun creating the machines.

My plan is to outline and explore the NIMP. I will compare it to other forms of the NIP and, in doing so, attempt to uncover the intuitions which indicate that it – and the problem of creative responsibility – is a genuine moral issue worthy of serious attention. I will then discuss possible solutions to the NIMP. This will involve illustrating the various ways in which the NIMP is more formidable than the standard NIP.

It is best to start, however, with the traditional NIP.

1. The Non-Identity Problem

1.1 Introducing the problem

The NIP concerns our obligations to future people. More specifically, it usually concerns obligations to future entities that are not themselves the product of human design choices (see Parfit's "14-Year-Old-Girl" case (1984, 358)). What do we owe to human beings who do not yet exist?

Consider an example. A country might face a choice: 1) commit to drastic conservation measures specified in, say, a global climate accord; or 2) pull out of the accord and allow fossil fuel corporations to do as they wish. Suppose that, on the preservation option, the quality of life for those living 100 years in the future would be substantially higher than on the laissez-faire alternative. Relative to the other option, either choice would have a large impact on the details of people's lives now. They would meet different people and, accordingly, have children with different people. As a result, current people would have different children in the different futures: the people living 100 years after the country chooses conservation are *different people* from those who would have lived if the country had abandoned the climate accord. So, do we benefit the people living in the future conserved climate? No, because if we had chosen differently, those people would not exist. Do we harm the people living in the future hellish climate? No, because if we had chosen differently, different people would exist.

This is the NIP. The people in the two possible futures are non-identical. On the "better" alternative, we do not benefit our progeny because they are not better off than they would have been otherwise. On the "worse" alternative, we do harm them because they are not worse off than they would have been. They simply would not have been. Hence, in terms of the future people, it does not matter morally which choice we make. A deeply counterintuitive conclusion!

The NIP is an argument. Arguments have premises. Therefore, the NIP has premises. Let us make explicit four of them. With one exception, each represents a possible solution to the NIP: by rejecting the premise, one can avoid the problem's conclusion. As a corollary, this also illustrates that the NIP is a problem faced by any moral theory so long as it maintains responsibilities to future individuals – though, of course, what form the responsibilities take, and what terms we would use to describe them, will vary by theory.[1]

1.2 Personal Identity

The NIP relies on the idea that there would be different future people in the different outcomes of our current choices. This is seen in Parfit's "Time-Dependence Claim," which suggests that who a person is depends on when he was conceived and who the parents are (Parfit 1984, 353).

Philosophers tend not to attack the NIP by questioning this view of personal identity. Parfit thinks that his claims are "easy to believe" and "not controversial" (Parfit 1984, 351). Nevertheless, to solve the problem through personal identity, we would need to claim that the future people would be the same in the different outcomes of our current choice. As a result, the current choice not to conserve the environment harms people living 100 years from now because, even if we had chosen differently, *the same people* would exist and have better lives. We are in fact making particular people worse off by our environmental negligence. Hence we are harming them. There is no non-identity and thus no problem. All choices are what Parfit calls "same people choices" (Parfit 1984, 356). (For a discussion of the different types of possible future worlds, see Carter 2001. For the view that merely possible people are morally irrelevant see Weinberg 2008.)

1.3 A life worth living

The notion of a "life worth living" is meant to restrict the number of cases that yield NIP puzzles. If, as a direct result of our current choices, a person in the future has a life not worth living, it is intuitive to say that we wrong the person. The wrong is established without claims about identity. We face the NIP only when the person *does* have a life worth living: they would not be better off not existing (because, by hypothesis, their life is worth living) and they would not be better off if we were to make different choices today (because then they would not exist). The notion of a life worth living, then, serves as a qualification on the NIP: *insofar as the future people have lives worth living*, how do we justify the idea that wrecking the environment today will harm them? This is how Parfit presents the problem (1984, 358–59). Hanser's discussion helpfully schematizes the role that a life worth living has in the argument (1990, 51, 59).

To have a life worth living means that the life meets some threshold of moral or existential acceptability, whatever that means. It is paradoxical because, as Epicurus asked, what exactly does a person gain by not existing? But insofar as the threshold is met, the NIP suggests that the unfortunate parts of the person's life cannot be called harms in the morally relevant sense. This is true because the worthwhileness of life is taken to have a morally compensatory effect (see Weinberg 2008, 4).

Unlike the other premises, rejecting the notion of a life worth living, or how it is employed in the premise, does not amount to a response to the NIP. In fact, since the notion serves to restrict NIP cases, rejecting it risks making the NIP even more imposing. I bring up the notion of a life worth living because it is central to the NIP and will enable us to think about the additional challenges of the NIMP.

1.4 Worse off principle

Key to the NIP is the idea that our current choices do not make future people *worse off* than they would have been if we had chosen differently. Identifying harms involves *comparing* counterfactual states of affairs. (To see the worse off principle in the original articulations of the NIP, see Kavka 1982, 95 and Parfit 1984, 374.)

Attacking the comparative theory of harm is the most common approach to solving the NIP. If a non-comparative theory of harm is true and relevant for the cases that the NIP trades on, the NIP dissolves. (Woollard (2012, 681–83) takes the view that both types of theories are morally relevant. Parfit and

Harman appear to take this view as well (Harman 2004, 109n12).) Accordingly, those hoping to solve the NIP through critiquing the worse off principle must show both that 1) there is a plausible non-comparative theory of harm and 2) the theory enables us to make sense of obligations to future generations. Parfit's idea is that the non-comparative sense of harm is not the morally relevant sense of harm in NIP cases (1984, 369; cf. Woollard 2012, 685).

Shiffrin (1999) sketches one prominent non-comparative theory. Harman (2004) offers another. I will provide the details of the non-comparative theories in 2.4. However, following Boonin (2008), whatever intuition-preserving power the non-comparative theories might have, they yield counterintuitive consequences of their own. If we are driven to non-comparative theories of harm out of a desire to preserve intuitions, we are driven away from them for the same reasons. The NIP literature can at times feel like philosophical whack-a-mole.

1.5 Person-affecting

The final premise I will consider concerns whether a harm must be a harm *to something*. For the NIP to work, benefits or harms must accrue to some particular individual. This idea is known as the "person-affecting principle" (see Parfit 1984, 394 and 396–401; Hare 2007, 499; Boonin 2008, 132 and 139. Korsgaard (2018) uses the term "tethering" to communicate the same idea).

The principle has two versions, "narrow" and "wide," though the labels are misleading because the difference between them is not one of scope or gradation. Better labels might be "token" and "type," respectively. The *narrow* person-affecting principle deems an action wrong only when the harms accrue to a specific (token) individual. The NIP exists on the narrow version of the principle. The *wide* person-affecting principle deems an action wrong when harms accrue to a population, regardless of the identity of the individuals who comprise the population (see Weinberg 2004, 5).

One way to solve the NIP is to avoid the narrow version of the principle. On this approach, although we actually do not harm the particular individuals in the future, we can call an action wrong by saying that an action harms the population or by pointing out some other bad-making feature. For example, we should conserve the environment because the future conserved world has a population that is, as a whole, better off than the population in the laissez-faire alternative (see Harman 2004, 90).

1.6 The NIP and future machines

With the four premises we have a better picture of both the NIP and the available routes toward a solution. Different moral theories will go in different directions. Staying within the framing of narrow problems of creative responsibility, I will now shift to the context of future machines and from "when" choices to "how" choices. The following section will have the same structure as the first: I will introduce the problem and then discuss altered versions of the four premises. My aim is not to solve the NIMP here but to shed light on its distinctive challenges for the purpose of encouraging deeper thought on the issues of public policy that inform our design choices in technology. We must not focus only on existential risk/safety and the moral machine problem.

2. The Non-Identical Machines Problem

2.1 Introducing the problem

Consider an example. Imagine a god contemplating the future of its creation. It decides to populate an Earth-like planet with some entities. We can conceive of a vast space of possible types of entities. The god is facing a choice: which entities should it make? We are familiar with human beings, as diverse as

they are. They occupy a point or cluster of points in the space of possible types of entities. We could conceive of something different. The god could create a creature that feels excruciating pain with every step it takes, struggles to think through problems, routinely fails to predict the consequences of actions, and is easily prone to fits of rage. With the whole space of possibility available, if the god makes those miserable creatures, does the god wrong its creation?

We can conceive of another possible created entity. The god could create a creature with senses finely attuned to pleasure, a body that is impervious to painful diseases and injuries, a level of intelligence that enables it to reason through abstract problems and discover truths about its environment with prolonged and enjoyable deep focus, a level of emotional and social intelligence that leads to political harmony, and no body odor. It is in the god's power to create these creatures. If it instead creates something like human beings or the miserable creatures, has the god wronged its creation?

Consider another case. The god itself is an entity, and thus it exists as a possible type of entity. From the god's perspective, it might judge that it has flaws that it can "fix" vicariously in its creative act. Maybe the god wishes to make something with superior memory, more emotional stability, or greater speed. There is a host of issues here. First, how reliable is the god's judgment of its own flaws? Perhaps the fix will lead to what the god would deem to be more severe flaws. Second, supposing the god is inexperienced and not omniscient, how well can it predict the consequences of its actions? Third, is it permissible for the god to create on the basis of the creation's likelihood to make the god's existence better?

How do these cases illustrate the NIMP? In the case where the god creates a miserable creature, has the god harmed the creature in creating it? Our urge is to say yes.[2] But if the god had created the creature differently, the miserable creature would not exist at all. A different, non-identical creature would exist. With the whole space of possibility available to it, the god cannot benefit or harm its chosen creation. The entities are not better or worse off than they would have been on some different choice because, on a different choice, they would not exist. These cases are meant to draw out the intuition that the creator *in creating* has obligations to its creation. The act of creation has moral significance. We are inclined to say that the god, since it has so many options available to it, should create an entity that would or could have a good existence. It is difficult to answer the question of exactly which entity the god should create. It is easier to see that the choice has moral import for the creation.

There are limitations to the god analogy, of course. First, unlike a god, technology companies do not have the power to create their machines *ex nihilo*. This introduces an important level of complexity and uncertainty. What we choose are the techniques and priorities. We then speculate about the finished products. Second, the theological language downplays the reality that the technological "god," and the other gods living with it in the pantheon, will interact with their creation. The god lives, mortal and vulnerable, in the world among its creations. As it happens, the god has already made the choice to create something stronger, smarter, and faster than it.

We can now abandon the analogy and make the NIMP explicit. Many companies are working to create machines. Each machine exists in the space of possible types of entities. The companies' goals, and the techniques they use in hopes of achieving them, are the result of choices. Do the choices carry moral significance? Here we can insert the moral restrictions or obligations placed on the creator god in our earlier examples. We can imagine similar examples of machines that have miserable or pleasant existences. Are we obligated to make machines that fall within a morally-confined space of possible entities?[3] Is our deliberation over the choice constrained by moral considerations? It seems like it is in some way.

The NIMP concerns the morality of the choices that affect what type of currently non-existing, and likely unprecedented, entity will come to exist. Whether 1) the NIMP is different in kind from the NIP, 2) the NIMP is a type of NIP, or 3) the NIMP is a technologically-oriented NIP does not matter for my argument. Although, as I remarked, the "when"/"how" choice distinction marks two genuinely different moral components of a possible case, how we sketch the boundaries of the NIP with respect to the distinction does not particularly matter.

We can now turn to the premises. My main goal is to show that proposed solutions to the NIP cannot be easily imported to the NIMP. The truth of some of the NIMP premises might seem obvious, but I discuss them also to present implications that they have in the context of technology.

2.2 Personal identity

Regardless of how strong the claims about identity are in the NIP, they are far stronger and more perplexing in the NIMP. The attempt to solve the NIMP through personal identity is therefore more challenging. For Parfit, it was impossible for me to have had different parents. Time is relevant because my parents could have had their first born child at various times.[4] Although it is plausible to say that I could not have been born in the 1840s, there is nevertheless something compelling about such cases. We talk in these terms all the time. To say that I would not exist if I had been born in the 1840s strikes us as odd, at least grammatically. It is not obviously incoherent, even if it implies that I would have had different parents.

The NIMP involves a different identity claim and avoids the 1840s puzzle. Is it possible for you to consider the counterfactual of being born a chicken? If I had been born (or hatched?) a chicken, I would not have had my parents. I would have had chicken parents. Is this possible? Obviously not. I, a human, could not have been born a chicken. (Being *turned into* a chicken is a different case.)

The increased difficulty of the NIMP is seen in the fact that the identity claims are not focused on time- or origin-dependence, as they are in the NIP. Because we are considering the difference between two different *types* of entities, when they are conceived/created or who their parents are is irrelevant. If the god creates the miserable creature, then it creates an entity that is not identical to the body-odorless creature. This is all that must be accepted for the NIMP. The non-identity is derived simply and uncontroversially from the difference in types. This leaves little room for one to reject the NIMP-version of the personal identity premise.

A way to solve the NIP is to claim that personal identity is not morally relevant. For Kumar (2003), what matters in our actions are not particular individuals but "types," the relational positions we hold in an action (e.g. "parent" or "child"). This is similar to Hare's ingenious solution to the NIP that appeals to "de dicto goodness" (2007, 516–23). Narveson (1967) discusses a similar idea. Unfortunately, their solutions cannot be imported to the NIMP because the "types" or "*dicta*" are precisely what is in question. We might have obligations to a *de dicto* human child because we understand the place that human children as a type occupy in the moral community. Specifying the type and articulating our obligations to it are significant, but whether the type should exist – and with this, how the type should change through technological development – is precisely what is at stake in the NIMP.

There are difficulties in applying these considerations to machines. Much of the language around personal identity is ill-equipped for machine cases. A machine is, to say the least, unlikely to be biologically conceived. It is likely duplicable. Perhaps most importantly, it could undergo rapid, radical, and frequent changes to its underlying architecture. The *number of entities that it is* could be indeterminate or constantly in flux (see Vernor Vinge's story "Long Shot" (collected in Vinge 2002) for an exploration of this possibility). Simply put, the existence of such sophisticated machines will (likely) require a new or

expanded conception of personal identity. The fanciful thought experiments that populate the personal identity literature are slowly becoming realities. Even in a world of humanoid intelligent machines where similarity to human beings provides a touchstone, the list of personal identity problems remains lengthy, and there is an assortment of tempting anthropocentric biases.

In much the same spirit as this paper, DiGiovanna outlines a similar series of challenges (2017). Future technologies might present us with "para-persons": entities that exhibit both the characteristics of personhood and further characteristics that undermine personhood. DiGiovanna's main example is the ability of an entity to erase and rewrite itself quickly. Debates about personal identity are always closely tied to issues of moral responsibility. The NIP is an instance of the connection. DiGiovanna's underlying concern is that the issue of para-personhood is indicative of the general challenge that our canonical ethical theories, insofar as they draw on theories of personal identity, are at risk of not being viable in the future.

The concern can be interpreted in two ways – as about either the moral machine problem or the problem of creative responsibility. Forgive a speculative point on each. On the former, I worry that our attachment to traditional theories of personal identity, or our resistance to expanding them (either due to anthropocentric/substratist bias or because of the apparent philosophical intractability of such a task) will lead us to misdraw the boundaries of the moral community – that is, to give incorrect answers to the moral machine problem (Biondi 2018a). On the latter, my interest in solving the NIMP *as part of* a more general project of solving the problem of creative responsibility, including its broad construal, inclines me toward severing the link between personal identity and ethical theory. But that is an issue for another time.

There is an additional challenge to artificial identity. Because AI development is not *ex nihilo* creation, we are creating ever more sophisticated machines through particular techniques. With respect to personal identity, because of the uncertainty and speculation, we face an opaque and more challenging version of the NIMP. Our daily choices are contributing to and changing our future machines. It is difficult to know how these choices will affect their identity.

When we combine the realities of technological development with the intuitions drawn out by the NIMP, we find a crucial implication: the current techniques used in creating future machines have moral significance. If a god, as creator, has obligations to its creations, the god will face those (or similar) obligations even if it has to build its creation *ex materia*. The techniques used to create intelligent machines are homing in on a particular area in the space of possible types of entities. Our current design choices are limiting or constricting the space, making the existence of some types of machines more likely than others. Even if the current machines are not moral entities, the direction has moral implications. So the current practices in AI development must be scrutinized with the eventual, and perhaps the current, machines in mind, depending on the best theory of ethics and personal identity. The act of creating intelligent machines has already begun. If the act is constrained by obligations, the obligations are in effect now.

2.3 A life worth living

It is important to note the difference between the moral significance of life *itself* and a life that will almost certainly involve morally significant experiences of a certain kind. A common view is that existence is morally neutral. We see it voiced by Hare as the "moral neutrality of existence" (2007, 509; see also Weinberg 2007, 14). Fortunately, the move from life to existence is justified because the concept of "life," with its biological connotations, is not what is important, notwithstanding discussions of "artificial life" (Boden 1996; Swanson and Schwitzgebel 2018). A "life worth living" is shorthand for the unwieldy concept of a "morally or existentially acceptable quality of existence of a morally significant entity." So

far the entities concerned have just so happened to be/have organic lives; they've been entities whose lives have involved them being alive. I bring this up because we are facing the prospect of morally significant entities that are not alive in the sense applicable to organic creatures. We can instead consider the broader concept of an "existence worth having."

With the terminological fix in place, what complications do we face in the NIMP? If an entity is one of a kind, or of an unprecedented kind, how could we determine whether it has an existence worth having? It might answer if we asked, but how would it be able to make the determination for itself? How could we evaluate the answer? These are mysterious questions and there are numerous ways to interpret them. To help, we can mark an exhaustive distinction among possible types of entities: relative to an environment, there are (1) types of entities whose existence is *always* worth having, (2) types of entities whose existence is *never* worth having, and (3) types of entities whose existence is *indeterminate* with respect to its worthwhileness. Within (3) we can also inquire into the likelihood that the existence of the entities will be worthwhile. When a god is contemplating its creative choices, is it obligated to select entities that fall within (1), assuming there could be such types of entities? Can the god opt for (3) as long as there is a sufficiently high chance that the existence of the entities will be worth having (or a sufficiently high percentage of the entities it creates will have existences worth having)? There are a number of problems here.

Certainly a god should not create entities that for some reason *cannot* have existences worth having. Perhaps the miserable creatures discussed earlier in this paper fall into category (2). In that case, as with the NIP, we can restrict the NIMP to cases in which the future machines have existences worth having. The NIMP would run as follows: *so long as the future machine has an existence worth having*, it would not be better off not existing (because, by hypothesis, its existence is worth having) and it would not be better off if we made different choices today (because then it would not exist). But how can we determine, prior to creating, where an entity belongs with respect to the three categories set out in the previous paragraph? Because we have never created an AI before, we have no reliable way to assess the quality of its existence. This is true not only because the AIs have not existed before but because they will likely be quite different from human beings. We must therefore solve two problems, one technical and one moral: we must determine what the existence of the machines will be like, from the perspective of the machines, before they exist (because, after they exist, the issue of whether they should begin to exist is moot) and also develop the moral theory that can determine whether such a machine should exist. On the standard NIP, we don't face the first problem, and thus the second is relatively more straightforward.

The NIP usually assumes that humans are the type(s) of entities that fall within (3). We can debate the categorization, but in the NIMP, the issue is broader. The god's creative choice is a choice among categories (1), (2), and (3). The space of possible entities therefore opens up before us. The concept of a "life worth living" or "existence worth having" is no longer a conceptual oddity, or a mere qualification on NIMP cases, but a notion that must be concrete enough to guide creative acts. Thus there are two levels to the NIMP challenge: not only must our analysis of the concept of an "existence worth having" be sufficiently intelligible to guide our choices, but it must expand to encompass more of the space of possible types of entities.

Now we stack on top of this the complexities of *ex materia* creation. In point of fact, we are already diligently creating the machines of the future without considering whether their existences are worth having. This might be a moral failure in itself.

2.4 Worse off principle

Before turning to theories that reject the worse off principle, we should linger on the complexities that the principle yields in the NIMP. It will be relevant to possible solutions. When we consider a human in a

wrecked environment and another in a pleasant environment, there is a sense in which the comparison procedure is perfectly straightforward. There are two possible human beings, and the future where one of them exists in a pleasant environment is preferable.

On the NIMP we are not comparing the well-being of two entities of the same type. But even with this complication, there is little reason to think that a comparison is impossible. Instead of the well-being of the same type of entity, we can simply consider the well being of any and all types of entities – humans, chickens, machines, and so on. Here we face the well-known problem of formulating a generalized comparison procedure that places the experiences of all entities along the same axis of value. Can we say that a particular human existence is better than a particular chicken existence? If we use pain and pleasure as the metric, what do we make of a sophisticated AI that is not built that way? We could easily list more complications.

Nevertheless, even if we were to construct a metric by which to carry out such a comparison, there is yet another layer of difficulty. In the NIMP we are comparing types of machines that do not currently exist. And when one machine exists, we are comparing its well-being with that of many machines that have never existed. We are unlikely to have any precise idea of what they would be like. The procedure would be full of hazy speculation. Hence, if we endorse the worse off principle, the strengthened identity claim in the NIMP makes it more difficult to argue that a machine is harmed when we make it one way rather than another.

Perhaps we should reject the worse off principle and accept a non-comparative theory of harm. Shiffrin, for instance, articulates one: "Harm involves conditions that generate a significant chasm or conflict between one's will and one's experience, one's life more broadly understood, or one's circumstances" (Shiffrin 1999, 123). A reader might feel like this could easily slip into a comparison procedure. The "chasm or conflict" is between two things, one worse than the other. More precisely, the chasm is between one's experience and one's will – though the will is seemingly the will for a *better* (off) experience. But I set these issues aside. Shiffrin continues, "Typically, harm involves the imposition of a state or condition that directly or indirectly obstructs, prevents, frustrates, or undoes an agent's cognizant interaction with her circumstances and her efforts to fashion a life within them that is distinctively and authentically hers" (Shiffrin 1989, 123–24). On this account, harm involves being alienated from the conditions one would will. It prevents a person from achieving a harmony between their experience and their will.

Harman offers another theory. She states that "an action harms a person if the action causes pain, early death, bodily damage, or deformity to her, even if she would not have existed if the action had not been performed" (Harman 2004, 93). This is a non-comparative theory of harm, but it is also a (brazenly!) stipulated solution to the NIP. Harman is following Shiffrin in listing pains. The list of harms is "unified by *comparison* with a healthy state" (emphasis added).[5] So it is difficult to find where the non-comparative theory of harm actually is. Again it might feel like a comparison procedure is looming.

However, what if we set these worries aside and embrace a non-comparative theory? The NIMP is about the harm of creating a machine of one type rather than another. So we face an immediate challenge. The non-comparative theories are concerned with the harms of entities of a known and fixed type. They are not directly concerned with the harms of being an entity of a certain type. But maybe they could be. There are two points to consider. First, the issue of personal identity returns. There is something incoherent about the desire to be an entity of a different type. Despite the intelligible surface grammar, I cannot will to be a chicken. I either somehow think I already am a chicken in human form (like an otherkin individual) or I will for something that entails the end of my existence. (Being a human in chicken form is a different case.) These considerations might point us to the idea that non-comparative theories of harm only function when considering the different experiences of a particular entity with a consistent identity.

If so, they cannot help us make sense of the alleged harms in NIMP cases. Second, according to Harman and Shiffrin, the harms are pain, early death, bodily damage, and the like. Could the harm be simply an existence that is not ideal or sufficiently good? Is the existence of the miserable creatures a harm because of the pains or because of the existence itself? If the harm is the creation of one type of entity over another, it is difficult to see how the non-comparative theory can make sense of it. Of course the miserable creatures discussed in this paper would be benefitted, comparatively or non-comparatively, by no longer being in pain, but that is not the issue at play in the NIMP. We are interested in whether the miserable creatures should be created (even if we deem them to have lives worth living, or adjust the case accordingly) instead of other possible types of creatures.

For Shiffrin, harms interfere with the interests or will of an individual. Surely an entity need not recognize and take issue with its own existence for the act of its creation to be a wrong. If a god were to create the miserable creatures, the god might make their shortcomings invisible to them. This is a common theme in AI science fiction. It is intuitive to think that the creations have still been harmed. Even more so in such a case! But in these cases, the non-comparative theories might not currently be expansive enough to make sense of the alleged harms to future machines. We will also need to encounter the issue of what a "will" will be in these machines. But more fundamentally, and more to the point, should we create something that has a will at all?[6] If we can but do not, are we harming our creation? The non-comparative theory does not give us direction in these cases.

What, however, if we assume that a non-comparative theory of harm is viable and relevant in NIMP cases? Because the space of possible entities of future machines is vast, there are many ways for creators to go astray. The likelihood of harm rises as the space of possible entities expands. This leads to the pragmatic issue. We may find we have solved the NIMP by accepting a non-comparative theory as morally relevant, but we now need actionable insight. What motivates the NIMP is the practical problem of how we direct AI development through the space of possible types of entities. Even if a non-comparative theory enables us to make sense of harms to future machines, it is difficult to find concrete ways to avoid the harms. We must do more than diagnose.

2.5 Person-affecting principle

Here the goal is to avoid the narrow person-affecting principle. The principle, recall, states that harms must accrue to specific individuals. There are no free-floating harms that are not harms *to* somebody or something. (To avoid the issue of personhood, which is not directly relevant here, I phrase the principle more broadly.) Consider Parfit's Q claim, which can be read as the wide version: "If in either of two possible outcomes the same number of people would ever live, it would be worse if those who live are worse off, or have a lower quality of life, than those who would have lived" (Parfit 1984, 360). We can speak of a population being better off than another, even if the individuals in the populations are entirely different. The Q claim, it is important to note, is only about populations of the same size. Parfit argues that if the wide person-affecting principle is applied to different number choices, we are led to a series of repugnant, absurd, and ridiculous conclusions (Parfit 1984, chs. 17–20). I will not detail the conclusions here. I simply wish to mark again that particular approaches to solving the NIP, if successful, soon face further problems.

Another version of the wide person-affecting principle would be what Harman calls the Impersonal Explanation: "The correct explanation of the impermissibility of the action is not that it harms: it does not harm. Rather, the action is impermissible because the world is better if the action is not performed; it is impersonally better, though it is not better for any person" (Harman 2004, 90). For example, we should conserve the environment because the future conserved world is impersonally better than the laissez-faire alternative. Although we do not harm particular individuals in our choice, the aggregate well-being of the

different populations will determine which choice we should make. In this case, as Harman's phrasing appears to suggest, we might not be talking about harms at all.

With the NIMP, the debate is different because our choices are not between two possible future populations of machines of the same type. That is, if the two possible future populations are of different types, we would struggle to apply something like Parfit's Q claim or Harman's Impersonal Explanation. The wide person-affecting principle is about different populations of the same type of entities. As we saw in the previous section, we might propose ways of broadening the principle or its application. This would require facing the challenge of constructing the universal metric. Further, not only would the two groups of machines be difficult to compare with respect to well-being (or whatever the criterion is), but one of the groups would have never existed. And, of course, there are far more than two groups at play. So if we want to use the wide person-affecting principle to solve the NIMP, we would need to adjust it and supplement it with confident counterfactual theorizing about possible alternative populations of different machine types. Both are daunting tasks.

As a final more speculative point, in the NIMP we would need to relate the person-affecting principle to an entity with uncertain or unprecedented personhood status. We face this problem when discussing the moral significance of non-human animals, fetuses, and DiGiovanna's "para-persons." But the problem is made more difficult by the fact that the machines are 1) the result of human choices, at least in part, and 2) created *ex materia*. On the first point, not only do we not know when a machine becomes a "person" we could "affect," but our design choices are determining what the entity is and becomes. Certain assumptions about personhood, or the traits relevant to it, are being brought to bear, at least implicitly, in the act of creating new entities. To some, this might appear to solve the moral problem: if we create something in our own image, we thereby know its personhood and moral status. And if the intelligent machines are different from us, we will debate the moral relevance of the differences by comparing the machines to ourselves. My point is different. As I have been arguing, the discussion should be about more than the moral machine problem. It is true that, before applying the person-affecting principle, we must ask whether we have built persons. But the discussion is also about the responsibilities that constrain creators: "which machines should we make?" not "what should we make of machines?" This is the core of the problem of creative responsibility. For the NIMP, personhood or moral status is not about the categorization of entities. The problem highlights and questions the assumptions that are reflected in our design choices, techniques, and visions of the future.

This is related to the point about *ex materia* creation. As I noted, I am motivated by the possibility that our moral obligations to future machines are in some sense active and binding now. One of the most significant aspects of the NIP is the idea that the morality of our current choices will be manifest in the future. This is a point of similarity between the NIMP and NIP. In the NIMP we find ourselves in the position of the god in our earlier examples, with the space of possible types of entities in front us. We have begun creating. Since our current and past creative choices in AI development have restricted the space, we have already made choices of moral significance. We are simply unaware of whether we should be praised or blamed.

2.6 Final observations

I have been focused here not only on the theoretical plausibility of possible solutions to the NIMP, but also on whether a solution can help confront the pragmatic issue of which direction we should take the design of our technologies. The NIMP is about "how" choices of narrow creative responsibility and thus references a space of possible types of entities. Our design choices involve selecting, directly and indirectly, entities from the space. Which choices should we make? If our moral theory enables us only to criticize the choices once they have been made, the theoretical solution is partial and ignores the pragmatic spirit of the NIMP.

To the philosophers who are not bothered by the NIMP, I recommend beginning work on the construction of moral guidelines for design choices that have potential harms to the created machines themselves as the central concern. I have yet to see any such guidelines. And I hope to have demonstrated that we need them.

Notes

1. Some people accept the problem's conclusion that we do not have obligations to future generations (see Schwartz 1978 and Heyd 1994).

2. See Adams 1972 for a discussion of the Leibnizian principle (likely of broad creative responsibility) that a perfectly good being must create the best world it can. I am focusing the NIMP not on the creation of worlds but on the creation of entities in an already existing world – i.e. the focus is narrow, not broad.

3. This is complicated by the fact that designers might not create an AI directly. They might instead create machines or automated processes that create an AI. There are strong reasons to think that this is the more likely future – the most persuasive being that it is an accurate description of the present.

4. The phrase "first born child" has a *de re* and *de dicto* interpretation (see Hare 2007, 514; Boonin 2008, 136; Weinberg 2008, 11).

5. The passage quoted is followed by the parenthetical, "though I haven't claimed that all harms meet this condition" (Harman 2004, 111n22).

6. This verges on the broad problem of creative responsibility.

Bibliography

Adams, Robert Merrihew. 1972. Must God create the best? *Philosophical Review* 81: 317–32.

———. 1979. Existence, self-interest, and the problem of evil. *Noûs* 13: 53–65.

Benatar, David. 2008. *Better never to have been: The harm of coming into existence*. New York: Oxford University Press.

Biondi, Zachary. 2018a. The coming oppression. The Vim Blog. January 14. https://thevimblog.com/2018/01/14/the-coming-oppression/ (accessed November 18, 2019).

———. 2018b. Stop talking about the moral status of machines. The Vim Blog. December 22. https://thevimblog.com/2018/12/22/moral-status-of-machines/ (accessed November 18, 2019).

Boden, Margaret A., ed. 1996. *The philosophy of artificial life.* Oxford: Oxford: University Press.

Boonin, David. 2008. How to solve the non-identity problem. *Public Affairs Quarterly* 22: 129–59.

Carter, Alan. 2001. Can we harm future people? *Environmental Values* 10: 429–54.

DiGiovanna, James. 2017. Artificial identity. In *Robot ethics. 2.0*: *From autonomous cars to artificial intelligence*, ed. Patrick Lin, Ryan Jenkins, and Keith Abney, 307–21. New York: Oxford University Press.

Hanser, Matthew. 1990. Harming future people. *Philosophy and Public Affairs* 19: 47–70.

Hare, Casper. 2007. Voices from another world: Must we respect the interests of people who do not, and will never, exist? *Ethics* 117: 498–523.

Harman, Elizabeth. 2004. Can we harm and benefit in creating? *Philosophical Perspectives* 18: 89–113.

Heyd, David. 1994. *Genethics: Moral issues in the creation of people*. Berkeley, CA: University of California Press.

Kavka, Gregory. 1982. The paradox of future individuals. *Philosophy and Public Affairs* 11: 93–112.

Korsgaard, Christine. 2018. *Fellow creatures: Our obligations to the other animals*. New York: Oxford University Press.

Kumar, Rahul. 2003. Who can be wronged? *Philosophy and Public Affairs* 31: 99–118.

Narveson, Jan. 1967. Utilitarianism and new generations. *Mind* 76: 62–72.

Parfit, Derek. 1976. On doing the best for our children. In *Ethics and population*, ed. Michael D. Bayles, 100–115. Cambridge, MA: Schenkman Publishing.

———. 1984. *Reasons and persons*. Oxford: Oxford University Press.

Savulescu, Julian, and Guy Kahane. 2009. The moral obligation to create children with the best chance of the best life. *Bioethics* 23: 274–90.

Schwartz, Thomas. 1978. Obligations to posterity. In *Obligations to future generations*, ed. R.I. Sikora and Brian Barry, 3–13. Philadelphia: Temple University Press.

Shiffrin, Seana. 1999. Wrongful life, procreative responsibility, and the significance of harm. *Legal Theory* 5: 117–48.

Swanson, Will, and Eric Schwitzgebel. 2018. Is C-3PO alive? The Splintered Mind (blog). May 11. https://schwitzsplinters.blogspot.com/2018/05/is-c-3po-alive.html (accessed November 18, 2019).

Vinge, Vernor. 2002. *The collected stories of Vernor Vinge*. New York: Orb Books.

Wallach, Wendell, and Colin Allen. 2009. *Moral machines: Teaching robots right from wrong*. New York: Oxford University Press.

Weinberg, Rivka. 2008. Identifying and dissolving the non-identity problem. *Philosophical Studies* 137: 3–18.

Woollard, Fiona. 2012. Have we solved the non-identity problem? *Ethical Theory and Moral Practice* 15: 677–90.