Moral Uncertainty and Our Relationships with Unknown Minds

By John Danaher,
University of Galway

[accepted preprint of forthcoming article in *Cambridge Quarterly of Healthcare Ethics*]

Abstract: We are sometimes unsure of the moral status of our relationships with other entities. Recent case studies in this uncertainty include our relationships with artificial agents (robots, assistant AI etc), animals, and patients with 'locked in' syndrome. Do these entities have basic moral standing? Could they count as true friends or intimate partners? What should we do when we do not know the answer to these questions? An influential line of reasoning suggests that, in such cases of moral uncertainty, we need meta-moral decision rules that allow us to either minimise the risks of moral wrongdoing or improve the choiceworthiness of our actions. One particular argument adopted in this literature is the 'risk asymmetry argument', which claims that the risks associated with accepting or rejecting some moral facts may be sufficiently asymmetrical as to warrant favouring a particular practical resolution of this uncertainty. Focusing on the case study of artificial beings, this paper argues that taking potential risk asymmetries seriously can help to resolve disputes about the status of human-AI relationships, at least in practical terms (philosophical debates will, no doubt, continue), however, the resolution depends on a proper, empirically-grounded assessment of the risks involved. Being sceptical about basic moral status, but more open to the possibility of meaningful relationships with such entities, may be the most sensible approach to take, though this in turn creates tension in our moral views that requires additional resolution.

1. Introduction

In June 2022, Blake Lemoine, a Google-based AI scientist and ethicist, achieved brief notoriety for claiming, apparently in earnest, that a Google AI programme called LamDA may have attained sentience. Lemoine quickly faced ridicule and ostracization. He was suspended from work and, ultimately, fired. What had convinced him that LamDA might be sentient? In support of his case, Lemoine released snippets of conversations he had had with LamDA. Its verbal fluency and dexterity was impressive. It appeared to understand the questions it was being asked. It claimed to have a sense of self and personhood, to have fears, hopes and desires, just

like a human. Critics were quick to point out that Lemoine was being tricked. LamDA was just a very sophisticated predictive text engine, trained on human language databases. It was good at faking human responses; there was no underlying mind or sentience behind it.[1]

Whatever the merits of Lemoine's claims about LamDA, his story illustrates an ethical-epistemic challenge we all face: how should we understand our relationships with uncertain or contested minds? In other words, if we have an entity that appears to display mind-like properties or behaviours, but we are unsure whether it truly possesses a mind, how should we engage with it? Should we treat it 'as if' it has a mind? Could we pursue deeper relationships with it, perhaps friendship or love? This is an epistemic challenge because in these cases we have some difficulty in accessing evidence or information that can confirm, definitively, whether the entity has a mind. It is an ethical challenge because our classification of the other entity – our decision as to whether it has a mind or not – has ethical consequences. At a minimum, it can be used to determine whether the entity has basic moral standing or status. It can also be used to determine the kinds of value we can realise in our interactions with it.

Our relationships with AI and robots are but one example of a situation in which we face this challenge. We also face it with humans whose minds are fading (e.g. those undergoing some cognitive decline) or difficult to access (those with 'locked-in' syndrome). And we face it with animals, both wild and domestic. Our default assumptions vary across each of these cases. Many people are willing to presume that humans, whatever the evidence might suggest, have minds and that their basic moral status is unaffected by our epistemic difficulties in accessing those minds. They are less willing to presume that the value of the relationships they have are unaffected by these epistemic difficulties. Some people are willing to presume that animals have minds, at least to some degree, and that they deserve some moral consideration as a result. Many of them are also willing to pursue meaningful relationships with animals, particularly pets. Finally, most people, as of right now, tend to be sceptical about the claims that AI or Robots (what I will call 'artificial beings' for the remainder of this article) could have minds. This is clear from the reaction to Blake Lemoine's suggestions about LamDA.

In this article, I want to consider, systematically, what our normative response to uncertain minds should be. For illustrative purposes, I will focus on the case study of artificial beings, but what I have to say should have broader significance. I will make three main arguments. First, that the

correct way to approach our moral relationships with uncertain minds is to use a 'risk asymmetry' framework of analysis. This is a style of analysis that is popular in the debate about moral uncertainty, and has obvious applications here. Second, deploying that argumentative framework, I will suggest that we may have good reason to be sceptical of claims about the moral status of artificial beings. More precisely, I will argue that the risks of over-inclusivity when it comes to moral status may outweigh the risk of under-inclusivity. Third, and somewhat contrary to the previous argument, I will suggest that we should, perhaps, be more open to the idea of pursuing meaningful relationships with artificial beings – that the risks of relationship exclusion, at least marginally, outweigh those of inclusion.

In deploying the risk asymmetry framework to resolve the ethical-epistemic challenge, I do not claim any novelty. Other authors have applied it to debates about uncertain minds, before. In the remainder of this article, I will reference the work of four authors, in particular, - Erica Neely,[2] Jeff Sebo,[3] Nicholas Agar,[4] and Eric Schwitzgebel[5] – each of whom has employed a variation of this argument when trying to determine the moral status of unknown minds. The novelty in my analysis, such as it is, comes from the attempt to use empirical data and psychological research to determine the risks of discounting or overcounting uncertain minds. My assessment of this evidence leads me to endorse conclusions that are different from those usually endorsed in this debate (though, I should say, similar to the conclusions reached by Nicholas Agar). The other contribution I hope to make is to be more systematic and formal in my presentation of the risk asymmetry framework. In other words, irrespective of the conclusions I reach, I hope to demonstrate a useful method of analysis that can be applied to other debates about uncertain minds.

2. Understanding the Ethical-Epistemic Challenge

The ethical-epistemic challenge stems from (a) the epistemic inaccessibility of other minds and (b) the importance of having a mind to many of our ethical beliefs and practices. Let us consider both issues in turn.

Although there is disagreement on exactly what it is that constitutes the mind, it is generally agreed that minds are constituted by properties such as intentionality (aboutness), the possession of intentional states (beliefs, desires, intentions), sentience (capacity to feel pain/pleasure), consciousness, self-awareness, sense of self as a continuing and enduring

'person' over time, and so on. For present purposes, I follow the work of Daniel Dennett, and others, in presuming that having a mind is not an all-or-none thing. Different entities can have different degrees of mindedness.[6] Some animals, for instance, may have intentionality and sentience, without having self-consciousness or an enduring sense of self.

The basic problem with assessing the mindedness of another entity is that mental properties are private. The only person that truly knows whether you have the capacity to feel pain, or whether you have intentional states, is you. Third parties can only ever infer your mindedness from your external behaviour. Thus, they might presume that you can feel pain because you draw your hand away from open flames, or presume that you have a desire for ice-cream because you say you do and you consume the ice-cream with relish. In a sense then, all minds, apart from our own, are uncertain. It is possible – following the classic thought experiment concerning philosophical zombies – that other people behave 'as if' they have minds when, in fact, they do not.[7]

Most people, however, do not take such radical philosophical scepticism seriously in their day-to-day lives. They are happy to assume that other humans have minds, possibly based on reasoning from analogy: If I have a mind, they are likely to have minds too. They are also happy to assume that many animals have minds (to at least some degree) based on how they behave and interact with us (and perhaps because we share common biological origins and underlying mechanisms of behaviour). Indeed, if anything, people tend to *over-ascribe* mindedness in their day-to-day lives. Psychologists have long noted that humans – and other animals – have 'hyperactive' agency detection or hyperactive anthropomorphism: a tendency to assume that events they encounter in their natural environments are the products of mental action, even when they are not.[8] This may lie at the root of many primitive, animistic religious beliefs. Still, when we move beyond the human realm, we often have more reason to doubt the mindedness of the entities we encounter and, consequently, our uncertainty grows. This seems particularly true of our encounters with artificial beings, who are often designed to intentionally mimic human behaviour and intelligence, and trigger anthropomorphic responses.

The challenge of assessing mindedness is acute because it is central to many of our ethical beliefs and practices. Whether or not an entity has moral status – i.e. whether it ought to be an object of moral concern and possess moral rights relative to us – is often taken to hinge on whether it has certain mental properties, such as sentience or enduring self-

awareness. There are, to be sure, other theories of moral status. Some people, for instance, argue that certain features of the natural environment have moral standing,[9] even if they lack mentality, but even if they are correct, most people agree, at a minimum, that having certain mental properties is *sufficient* for moral status, even if it is not necessary. Furthermore, valuable human relationships – friendship and love – are often thought to hinge on the mental properties of the parties to those relationships. The Aristotelian ideal of friendship, for instance, hinges on an assumption mutual well-wishing: another person is not really your friend unless they care about you and wish you well. Likewise, most accounts of loving relationships presume that there must be mutual goodwill, voluntarily chosen commitment, and ongoing affection/enjoyment on both sides of the relationship. All of these things depend on mindedness.

A lot has been written in recent years about the possibilities of AI or robots having moral standing and being our friends or lovers. It is impossible to do full justice to this literature here, but some examples are provided in the endnotes.[10] As David Gunkel and Mark Coeckelbergh noted some time ago, a lot of this literature focuses on the mental *properties* that an artificial being would need to have in order to count as a true moral patient or a friend.[11] But given the epistemic inaccessibility of the mental, and the reasons to doubt the mentality of others, at the heart of this literature is an ineliminable moral uncertainty regarding the status of artificial beings and our relationships with them.[12] Some people have suggested that we can resolve this uncertainty, at least when it comes to basic moral standing, in favour of being ethically over-inclusive when it comes to artificial beings.[13] Others have suggested that there can be no universal resolution of the uncertainty, it all hinges on an individual's private subjective assessment of the probability that an artificial being has a mind.[14] Is this right? To answer that, we can draw from lessons learned in the debate about moral uncertainty.

3. Risk Asymmetry Arguments in General

There is a school of thought that argues that we need to take moral uncertainty seriously in our decision-making and that moral uncertainty, if left unresolved, can undermine our moral agency.[15] The core idea uniting this school of thought is that our moral beliefs are open to some doubt and, depending on which beliefs we endorse, we can reach radically different conclusions about what we ought to do.

Consider, for example, the distinct ethical guidance offered by consequentialist and non-consequentialist normative theories. Following a simplistic form of consequentialism, might lead one to conclude that it is morally permissible to sacrifice one life to save five (in the classic trolley dilemma); following non-consequentialism might lead one to the opposite conclusion. While some people might be 100% sure that consequentialism is the correct moral theory, most people have doubts. Their subjective credence for consequentialism is likely to fall well shy of 100%. The challenge for moral uncertaintists is to figure out what we should do, given this doubt. In response, moral uncertaintists have developed a range of meta-normative decision rules – e.g. maximising expected choice-worthiness[16] – that provide practical guidance for resolving our uncertainty.

Moral uncertainty comes in several forms, and the dividing line between it and factual or predictive uncertainty can be blurry. The most straightforward kind of moral uncertainty arises at the level of normative theory or principle ("you should do x" versus "you should do not-x"). But uncertainty regarding moral status or standing is also a kind of moral uncertainty. Here, we are uncertain as to whether an entity has the properties required to render it worthy of moral respect. This confronts us with a classic false-positive/true-positive dilemma. Consider uncertainty regarding the moral status of a person with locked-in syndrome (call them 'X'). We can choose to believe that they have moral status or not. Depending on what the actual status of X is, this results in four possible outcomes:

- True positive – X has moral status and we treat X as if it had moral status
- False positive – X does not have moral status but we treat X as if it had
- True negative – X does not have moral status and we treat X as if it does not
- False negative – X does have moral status but we treat X as if it does not

In table format:

	Act as if X has moral status	Act as if X does not have moral status
Has moral status	**True positive**: appropriate moral inclusion	**False negative**: inappropriate moral exclusion

Does not have moral status	**False positive**: inappropriate moral inclusion	**True negative**: appropriate moral exclusion

One influential line of reasoning in the uncertainty debate claims that, in some cases, the risks or harms associated with the different errors (false negatives vs false positives) is so asymmetrical as to warrant a particular practical resolution of the moral uncertainty. For example, there are risk asymmetry arguments about the moral status of animals[17] and foetuses.[18] The claim in both instances is that the risk of a making false negative error (assuming animals/foetuses lack moral status when they do not) is much higher than the risk of a false positive error (assuming that they do not have moral status when they do). The former results in continuous, impermissible and cruel treatment of something that deserves appropriate moral consideration; the latter does not. Consequently, we ought to act as if animals and foetuses have basic moral standing, as long as the uncertainty applies.

Set to one side, for now, the question as to whether these arguments about animals and foetuses are correct. For now, the challenge is to extract the logic underlying the standard risk asymmetry argument. It appears to be something like this:

- (1) You are faced with a choice between Option A and Option B but are unsure which of those two options is morally acceptable (permissible, obligatory etc).
- (2) In order to know which of the two options is acceptable, you would need to know moral fact X (which you do not know, or cannot know, at least not for sure).
- (3) If X were true, then Option A would clearly be morally unacceptable (high moral cost, low moral benefit) and Option B would be acceptable (morally neutral or some moral benefit).
- (4) If X were false, then both A and B would be acceptable but the relative benefits of A, if any, would be minimal.
- (5) You ought try to minimise expected moral costs (at least where those costs are obvious)
- 6) Therefore, despite your uncertainty with respect to X, you ought to choose Option B over Option A.

The goal of the remainder of this article will be to assess how well this argument applies to debates about the moral standing of artificial beings

and the potential for meaningful/valuable relationships with them. Before doing that, however, it is worth briefly reviewing some general criticisms of risk asymmetry arguments.

An obvious first criticism concerns the threshold of uncertainty required before the argument takes effect. Some have expressed concerns about infinitesimal uncertainty/risk. Maybe there is some really small chance that rocks are conscious and so, every time we stand on them, we cause them tremendous pain. Should we take such uncertainty seriously and factor it into our decision making? If the risk asymmetry logic is to be followed to its hilt, then maybe we should, but this then seems to reduce the logic to the absurd.[19] To take moral uncertainty seriously would be to undermine moral agency. Proponents of risk asymmetry arguments have responses to this criticism. Dan Moller, for instance, argues that we should simply ignore infinitesimal risks, as we do in other decision-making contexts.[20] Jonathan Matheson similarly argues that we should focus on significant risks and not marginal ones.[21]

But this response is not entirely satisfactory. What counts as a sufficiently significant risk? Some people think the chances of a robot being conscious or caring about you are vanishingly small. For them, the risk asymmetry arguments I am going to consider in a moment have no appeal. One possible answer here is to claim that the appeal of risk asymmetry arguments will vary from person to person. Your own subjective probabilities regarding the moral status of robots (or whatever is in moral doubt) determines whether you take the argument seriously. If you have significant doubts, you will take the argument onboard; if you do not, you will ignore it. Nicholas Agar adopts essentially this stance when he deploys the argument in the debate about artificial beings.[22] Another possibility is that whether the argument reduces to the absurd depends, crucially, on the actual moral costs and benefits involved in any particular decision. It might sound absurd to take the moral standing of rocks seriously, but that is arguably only because people raising this objection assume that if we did take it seriously, the moral risks of mistreating rocks would greatly outweigh the moral benefits of continuing to treat them much as we currently do. That might not be true. The uncertainty might be sufficiently small, and the moral benefits of continuing with business as normal so high, as to warrant continuing with the status quo. In other words, we should not discount the risk asymmetry argument so quickly, without doing some proper accounting of the relative costs and benefits of false positive versus false negative errors. That is one of the things I attempt to do in what follows.

There are other criticisms of risk asymmetry arguments. Do they necessarily presume a consequentialist approach to moral decision-making? Not quite, though the argument sketched above (we ought to minimise moral costs) does lean into consequentialism at the meta-normative level. If, for some reason, you think this approach to decision-making is fatally flawed or out of bounds, the argument may have no appeal. Other objections are quite technical and abstruse. Brian Weatherson, for example, has argued that proponents of such arguments make an error in how they approach morality.[23] They value doing the right thing *de dicto* and not, as they should, *de re*. To put it another way, they fetishise doing the right thing, whatever that happens to be, and do not focus appropriate attention on the moral reasons for and against particular actions. There is not time and space to evaluate Weatherson's argument at the length it deserves (see Matheson for a critique).[24] Still, it seems like the risk asymmetry argument does capture something plausible about the nature of reasoning under conditions of uncertainty. Variations on it are commonly deployed by decision theorists in other contexts. It is hard to see why moral uncertainty should be treated so differently from other kinds of uncertainty. For the remainder of this article, I will presume that the risk asymmetry analysis has validity and consider its application to debates about the moral status of artificial beings and the value of our relationships with them.

4. Basic Moral Standing and Risk Asymmetry

Let us consider uncertainty about the basic moral status of artificial beings and the potential risk asymmetry that can be generated as a result. First, it is important to clarify what it means to recognise that another entity has basic moral status. For present purposes, I assume that to recognise basic moral status means recognising that another entity has moral rights and interests of its own, and that it is not simply an object or tool that we can use for our own purposes. Put another way, it is to recognise another as an object of moral reason and not simply practical reason. It is to see it as a member of our extended moral community and not just a thing in the world.

That said, 'basic moral status' may be a misleading phrase. It is likely that moral status comes in degrees. A mouse may have some moral status but maybe not the same moral status as an adult human being. The different degrees of moral status make the application of risk asymmetry arguments

complex. You might need to run a different asymmetry argument depending on the presumed relative degree of moral status. For example, the relative cost of falsely assuming that a mouse has a small degree of moral status is, perhaps, very different from falsely assuming that it has a high degree of moral status (i.e. similar to that of a human being). Assuming the former might only entail very minor changes in our current moral beliefs and practices – perhaps a little bit more care and attention to how we treat our rodent brethren. Assuming the latter might require a more substantive rearrangement of our moral beliefs and practices – eliminating all pest control measures, dedicating more time and attention to protecting mouse interests and welfare. If mice deserve such full recognition, then this substantive rearrangement is warranted, but if they do not, then the costs, in terms of wasted time, attention and resources, might be significant. To put it more abstractly: all choices have opportunity costs; the opportunity costs of giving mice more recognition than they truly deserve could be high, if you presume a high form of moral status, but low if you presume a low form.

For present purposes, I will assume that, in debating the moral status of artificial beings, we are concerned with whether they or not they have (or might have) a relatively high degree of moral status. Perhaps this is not fully equivalent to that of a human being, but it might be pretty close. So, at a minimum, it would imply that the artificial being has a right to exist/a right to life and a right not to be harmed or treated cruelly. I will run the risk asymmetry analysis with this in mind. Readers can adjust the analysis provided below depending on the supposed degree of moral status. In doing so, they will probably find that my assessment is much less persuasive if we only suppose that artificial beings are being granted a very low degree of moral status.

In running the risk asymmetry analysis we have to consider the costs and benefits of false positives (granting them moral status when they do not deserve it) versus the costs and benefits of false negatives (denying them moral status when they do). It is worth emphasising that we need to focus on costs and benefits in both cases. This is something that might be lost in the typical presentation of a risk asymmetry argument, where the focus is entirely on potential 'risks' which usually translates to 'costs of error'. It is possible that false positive errors have both costs and benefits and it is possible that false negative errors have costs and benefits. To only focus on potential costs would be to conduct an incomplete risk asymmetry analysis.

So let us take it in turn. What are the costs and benefits of making a false negative error? I start with this because many previous writers have suggested that the false negative costs are sufficiently high as to outweigh the false positive ones. This is the view of Erica Neely, for instance, when arguing about robots; it is also the view of Jeff Sebo and Dan Moller when arguing about animals and foetuses, respectively.[25] What is the reasoning here? It is that if we continue to falsely exclude artificial beings from the circle of moral concern, then we risk perpetuating a moral catastrophe: we deny rights to those that deserve them, we treat them cruelly and cause them great suffering. Analogies with past moral crimes can sometimes be used to bolster this reasoning.[26] For example, many cultures, historically, denied at least some elements of moral status to slaves and women and, thereby, caused them great suffering and harm. We should not risk making the same mistake again. Or so the argument goes. Against this, are there any potential benefits to false negative errors? Perhaps. By denying moral status to artificial beings we may retain resources for ourselves and others that would otherwise go to them. However, this does not seem like a moral benefit. It seems quite self-interested and immoral, if anything. This is what leads people to think the costs of making a false negative error are very high – high enough to outweigh any potential costs of making a false positive error.

But before concluding that this analysis is correct, we need to provide some systematic evaluation of the costs and benefits of making a false positive error. What are they? In terms of costs, if we give an entity more moral respect than it deserves, we risk wasting time, attention and resources. This an opportunity cost: we could have spent the time, attention and resources on other, more deserving, entities. This is a common view among critics of the idea of recognising the moral status of artificial beings.[27] Some might argue, however, that this is compensated for by an increase in moral compassion and empathy. In other words, if we care about artificial beings (even in error) we might start caring more about other entities that really deserve our moral respect. A rising moral tide of concern will lift all boats, so to speak. This is akin to the argument a parent might make when they have a second child: the addition of the second does not divide or dilute their love and affection for their first child. Love is not a finite resource that must be divided among worthy recipients. On the contrary, it is something that can grow and expand to accommodate more deserving souls. So perhaps the cost of making a false positive error is significantly reduced (maybe even entirely compensated for) by the benefits of doing so.

Speaking as a parent to two children, this argument has its appeal. I would like to think that my capacity for love and empathy has grown with the addition of a second child. If asked, I would say that I love them both equally: my love for the first has not been diluted by the addition of the second. In fact, my capacity for love and empathy has expanded as a result of having to care for two children. But is this really true? I may love them both equally, and hope to live up to this ideal, but the practical reality of day-to-day life suggests that having two children does dilute and divide your attention. You can claim, in principle, to love them both equally, and that your capacity for love has grown, but you cannot dedicate the same energy and resources to the first child anymore. You have to share it out. You might be able to compensate for this with other aspects of your life, e.g. less time spent at work or on personal hobbies, and this compensation may allow for some expansion in love, but there is still a limited number of hours in the day. Something has to give. You have to give up the focus on work or hobbies. Maybe that is an acceptable loss, but there is no getting around the fact that there are opportunity costs to the addition of more people to your circle of concern.

This is where the typical risk asymmetry analysis ends: with a mix of common sense and intuition-pumping. Is there any evidence to suggest that the false positive costs (and benefits) are as suggested? There is some. Empirical studies of the psychology of the 'moral circle' have become more common in recent years and some of these studies are suggestive.[28] In particular, a study conducted by Joshua Rottman and his colleagues is worthy of closer scrutiny. In this study, of approximately 1000 United States residents recruited by Amazon Mechanical Turk, Rottman and his colleagues set out to discover whether moral concern was a finite resource and, crucially, whether people that expanded their own personal circles of moral concern to include non-human animals and elements of the natural environment (called 'tree-huggers' in the paper) did so at the expense of marginalised/stigmatised human beings.[29] The findings of the study suggested that moral concern was indeed finite and that those that expanded their circles did so at the expense of humans.

To be more precise, the study asked people to assign different entities to a location within a circle of moral concern. These entities included family, friends, ethnically different people, parrots, rainforests and so on. The findings suggested that most people put family and friends in the innermost circles of concern and then assigned other human beings to different, more distant, locations within the circle. Nevertheless, one third of the people assessed put non-human animals and elements of nature

ahead of marginalised or stigmatised human beings. In other words, for these people, expanding the circle of moral concern to the non-human world came with a definite opportunity cost: it meant de-ranking or deprioritising humans in favour of nature/animals. Furthermore, for all people assessed, moral concern appeared to be a finite resource: the aggregate amount of moral worth assigned to everyone within the moral circle was the same, irrespective of whether someone was a tree-hugger or a human-lover.

This study lends credibility to the concerns raised by critics of AI rights.[30] Although the study was about the tradeoff between marginalised humans and nature/animals, one could imagine a similar tradeoff taking place between humans and artificial beings. There could, in other words, be some significant costs to making a false positive error of over-including artificial beings within our moral circles. It may lead people to deprioritise (worthy) humans in favour of (unworthy) machines, with possible impacts in terms of lost time, attention and resources dedicated to those worthy humans. If the artificial beings are assigned a high degree of moral status, the opportunity cost could be very high.

That said, this is just one study. It is important not to read too much into it. When interviewed about it, Rottman was keen to emphasise its limitations.[31] He and his colleagues did not attempt to recruit a representative sample. It is entirely possible that different cultures or groups could have different rankings or approaches to ranking membership of the moral circle. Furthermore, the study would need to be replicated and expanded to see how robust its findings are. Still, the findings are suggestive and they are a piece of evidence in support of the idea that false positive costs could be quite high while the alleged benefits (i.e. growing empathy and concern) are non-existent.

Would this invert the risk asymmetry analysis and suggest that the false positive costs of moral inclusion outweigh the false negative costs (and are not compensated for by any associated benefits)? Perhaps. We would need to see how robust the finding is. Nevertheless, one thing that might make such a reversal of the asymmetry more compelling would be if the false negative costs could be easily mitigated or avoided. Recall, that the false negative costs arise from the potential for creating artificial beings that genuinely deserve high moral status but are, incorrectly, denied this status. One way of avoiding this false negative cost would be to simply refrain from creating such artificial beings, i.e. create artificial beings that definitely do not have any moral status (or any potential risk of having such

status) or, more radically, refrain from creating them at all (in much the same way as we currently refrain from creating cloned humans). After all, the situation with respect to artificial beings is quite different from the situation with respect to non-human animals/nature or other marginalised humans. All these other entities either already exist or would continue to exist irrespective of our choices. With artificial beings we have a choice.

This is, in essence, the position advocated by some critics of AI rights, such as Joanna Bryson.[32] They say we should not create artificial beings that elicit potential moral sympathy or concern. We should keep them as tools, nothing more. However, this might be easier said than done. I do not subscribe to any strong form of technological determinism, but I am also not sure that we are entirely in control of our technological future. The temptation to create artificial beings is very strong, at least among some people, and the risk of creating ones with debateable moral status seems quite high (Blake Lemoine's doubts about LamDa already indicate this). If we cannot stop them from being created, then the costs of false negative errors cannot be eliminated.

The seeming inevitability of creating artificial beings with contested moral status is something that leads Eric Schwitzgebel to posit that we are on the cusp of a 'robot rights' catastrophe.[33] Either we will massively over-ascribe moral status to undeserving artificial beings (with the associated false positive costs) or massively under-ascribe moral status to deserving artificial beings (with the associated false negative costs). Either way, catastrophe awaits. This analysis, however, might be too fatalistic and, on the contrary, gives us an even stronger reason to try to avoid creating artificial beings. Perhaps we will not succeed, but if there is a chance that we could, or that we could at least reduce their number, we should because doing so would reduce the false negative costs.

In conclusion, the typical risk asymmetry analysis of moral status assumes that the false negative costs of under-inclusion (and false positive benefits of over-inclusion) massively and decisively outweigh the false positive costs of over-inclusion (and the false negative benefits of under-inclusion). However, the, limited, available empirical evidence suggests that things are not so straightforward. There are some significant moral costs associated with over-inclusion, and the alleged benefits of over-inclusion are negligible or open to doubt. Furthermore, it may be possible to mitigate the false negative costs of under-inclusion, simply by reducing, or trying really hard to reduce, the number of artificial beings with contestable moral status that come into existence. So the typical risk asymmetry analysis

might get things wrong: the asymmetry goes in the other direction, in favour of under-inclusivity not over-inclusivity.

5. Friendship and Risk Asymmetry

So much for basic moral status. What about the potential for valuable relationships with artificial beings? There has been much philosophical interest in the idea of morally significant relationships with artificial beings.[34] As noted already, these relationships raise similar uncertainty issues to those that arise in relation to basic moral status. There is no doubt that people can and do form genuine attachments to artificial beings. But do these attachments realise any genuine value? Can a robot be a true friend or a true lover? Or are they fake or 'counterfeit'? The answer to these questions is usually thought to hinge on whether the artificial being has properties that make it possible for it to care about us, to freely choose us, to wish us good will and so forth. Let us focus on friendship, primarily, with the assumption that much of the analysis can be transferred to other morally significant relationships.

At the outset, it is important to acknowledge that friendship is a complex relationship type and there are many different accounts of what it is and what it takes for someone to be a true friend. The classic Aristotelian analysis, for example, divides friendships into three main classes: pleasure, utility and virtue. There is no doubt that it is possible to attain a pleasure or utility friendship with an artificial being since neither of these friendships hinges on whether the 'friend' has mental properties that enable it to form a bond of mutual affection with you. Virtue friendships are different: they require genuine reciprocity and mutuality. There is, consequently, more doubt as to whether artificial beings can meet the conditions for such friendship and so moral uncertainty rears its head when we consider these kinds of friendship.[35] The uncertainty justifies a similar risk asymmetry analysis.[36] Do we err on the side of false positive friendship-inclusions (i.e. assuming an artificial being can be a genuine friend when, in fact, it cannot) or false negative friendship-exclusions (i.e. assuming the artificial being cannot be a genuine friend when, in fact, it can)?

A complication arises here because it is not clear whether virtue friendship is best understood as an 'all or none' thing. For present purposes, I will assume that Helen Ryland's account of friendship is broadly correct: much like moral status, there are 'degrees of friendship', with higher forms of friendship requiring more conditions to be satisfied than lower forms of

friendship, and usually requiring a higher degree of commitment and mutuality from the friends.[37] If the debate about friendship with artificial beings were solely about lower forms of friendship, then the false positive risks of over-inclusion might be quite minimal compared with the false negative risks of exclusion. So, once again, I will assume the debate is about whether a relatively high degree of friendship is attainable with artificial beings and I will try to assess what this entails from a risk asymmetry perspective.

Before proceeding to the risk analysis it might be worth pausing and asking whether there is an important difference between the debate about valuable relationships and the debate about moral status. One might argue that friendship is all in the eye of the beholder. If you think you are in a genuine friendship with someone, and you act as if you are, then you receive the benefits of this assumption irrespective of the deeper metaphysical reality. This makes the debate about friendship rather unlike the debate about moral status. In the latter case, it seems as though, if your assumptions do not match the moral reality, you can cause harm or suffering to others. There are high stakes to getting it wrong. Friendship appears to be a more private affair where one's subjective beliefs are king: getting it wrong costs you nothing.

There are two responses to this worry. The first is that if you are fully committed to believing in (or rejecting) the possibility of friendship with an artificial being, then perhaps a risk asymmetry analysis is irrelevant to you. You will not worry about any potential loss of friendship with an artificial being. But it is likely that some people will have their doubts. They will not be sure whether they should fully commit to such a friendship, or whether the friendship is genuine. We all, if we are honest, have these doubts about our own 'friends' from time to time. For these people, the risk asymmetry analysis has some value. The second response is that a large number of moral philosophers seem to think that there are genuine, objective, costs (and benefits) associated with getting the analysis of friendship right.[38] For them, the structure of this debate is almost identical to the structure of the debate about moral status.

So what are the costs and benefits of making a false positive error with respect to friendship with an artificial being? The costs are, once again, best understood in terms of opportunity costs: you pursue a friendship with a false artificial friend when you could have pursued a friendship with a genuine human friend. You substituted an inferior good (if it is, even, a good) for a superior one. This argument can be bolstered by empirical

evidence that suggests that we have limited friendship budgets. Robin Dunbar's famous 'number' – 150 – is allegedly the average upper limit on the number of friends it is possible for humans to have.[39] Dunbar claims that multiple, parallel streams of evidence support the idea that 150 (give or take) is the upper limit on possible friendships for humans. Some researchers reject this. Patrik Lindefors et al have argued that there is much more variation in friendship budgets than Dunbar's analysis suggests, at least based on Dunbar's original method for deriving this number.[40] Their estimates are lower than Dunbar's original one, but they also note that there are very wide confidence intervals associated with these estimates such that higher numbers of friends are possible. Dunbar has offered a robust critique of their analysis, claiming that it rests on flawed statistical measures and an incomplete assessment of the evidence.[41] It is not possible to evaluate this debate here but, given its widespread use and broad evidential base, we can assume, for the sake of argument, that Dunbar is roughly correct: there is an upper limit on the number of friends it is possible for humans to have and that upper limit is probably in and around 150 (with the number of possible close friends being even lower than this). Accepting this, does provide support for the idea that committing to a 'fake' friendship with an artificial being comes with a potential opportunity cost of a real friendship with a human. That said, the saliency of this opportunity cost will depend on how close someone is to exhausting their friendship budget. I can safely say that I am a long way away from the upper limit so I am not in any immediate danger.

Another, potential, false positive cost is the risk of betrayal within a friendship.[42] This is a risk associated with all human friendships and, ironically or paradoxically, may be one of the things that makes them so valuable: close human relationships are high risk-high reward.[43] It is because we are so deeply entangled with each other that we can achieve meaningful connection (love) but this can quickly change to mutual disaffection (hate). Betrayal arises when a friend that you thought was loyal to you and acting in a manner consistent with your interests and values does something that undermines those interests and values. A lover that cheats; a friend that gossips. These are classic examples. The risk of betrayal within human-artificial being relationships might, however, be higher than that within human-human relationships simply because of how artificial beings are likely to be constructed and controlled by corporations. These corporations may use artificial beings for covert surveillance and data-gathering and use the information gathered to manipulate or influence the 'friends' in some fashion. Betrayal is an ever present threat or possibility within such friendships.[44]

What about false positive benefits? There are some. As noted above, if you think you are in a valuable relationship with someone, it is possible that you will derive a lot of pleasure (at a minimum) and other kinds of benefit from that relationship irrespective of the deeper metaphysical truth. We will say more about these kinds of benefit in a moment.

Let us turn our attention then to the false negative costs and benefits. The obvious benefit of false negative exclusion is that you retain the time and opportunity to pursue friendships with human beings. You may lose out on friendships you could have had with artificial beings but there are plenty of human beings out there (over 8 billion on the most recent count) so this is not a major foregone opportunity. The value of human friendships is not in doubt and it is, consequently, a more desirable option (a safer bet) than pursuing a friendships with an artificial being. There are, nevertheless, significant false negative costs to factor into the equation. To return, briefly, to the work of Robin Dunbar, he claims that in addition to the intrinsic benefits of friendship – pleasure, sense of belonging, shared experiences etc – there are a variety of well-documented instrumental benefits.[45] People with more friends tend to live longer, suffer from less cognitive decline, and report better physical and mental health over the long run. What is more, we are living in the midst of a widely-reported crisis of loneliness.[46] Many people, perhaps most notably middle-aged men, have few, if any, friends. For instance, Vivek Murthy, a former US surgeon general, argued in 2018 that American society is undergoing an 'epidemic of loneliness'.[47] A 2018 Kaiser Family foundation study found that 1/5 people from the US and United Kingdom, as well as 1/10 from Japan, report feeling lonely most of the time.[48] It also found that loneliness correlated with negative mental, physical and financial outcomes. So, notwithstanding the abundance of people in the world, people are lonely. Turning our noses up at an opportunity for valuable friendships may not be the most sensible approach to take in the midst of this crisis of loneliness.

There does not appear to be a decisive risk asymmetry here. There are potential costs to false positive inclusion, but how compelling they are depends on a number of other factors. While we may have a limited friendship budgets, many people, if the data on the crisis of loneliness is to be believed, are nowhere close to exhausting their budgets. They could benefit from the addition of extra potential friends. Indeed, adding those potential friends to the pool could be even more compelling if it turns out that the people most inclined to form friendships with artificial beings are less likely to form friendships with human beings. In other words, if, for

them, there is no serious opportunity cost involved with adding artificial friends to the mix. Is there any evidence to suggest that this is true? Possibly. There is some empirical evidence to suggest that the people that are most likely to anthropomorphise animals and machines – and so the most likely to connect with them -- are those most inclined to be lonely and isolated.[49] Perhaps this is a compensation, but it may also be a stable character trait. For these people, opportunities for friendship with artificial beings might be the most workable solution to the problem of loneliness: they just find it hard, if not impossible, to form friendships with humans. Similarly, the false positive cost of betrayal, while certainly serious, could be mitigated by appropriate regulation of the design and control of artificial beings.

In sum, I do not think it is possible to make a compelling risk asymmetry argument with respect to friendship with artificial beings but, if I were to get off the fence, I would say that the benefits of false positive inclusion, combined with the costs of false negative exclusion, *marginally* outweigh the costs of false positive inclusion combined with the benefits of false negative exclusion. There is no decisive asymmetry; just a marginal one.

6. Conclusion: The Need for a Meta-Analysis

Nothing I have argued here is definitive or decisive. It cannot be when we are dealing with moral uncertainty. Still, I have concluded that we should err on the side of excluding artificial beings from the circle of moral concern (or, more properly, against the creation of such beings in the first place) and in favour of including them within the circle of friendship. There is, obviously, some tension between these conclusions. It is likely that in creating artificial beings that fall within the circle of friendship we would end up creating beings that fall within the moral circle (or, at least, raise legitimate doubts about their inclusion). We cannot do one without the other.

This raises the need for a meta-risk analysis. Do the benefits of potential friendship outweigh the costs of moral over-inclusion? Or vice versa? I am not sure what the answer to this is. On the whole, I think some anthropomorphic bias is justified. If the inclusion of potentially worthy artificial beings comes at the expense of definitely worthy human beings, then we should favour the worthy humans, all else being equal. We could compensate for the loss of potential friendships by exerting more efforts to help people find time and space for human connection, and overcome any

personal or psychological difficulties they have in forming human friendships.

Nevertheless, in some ways, I hope that the conclusions I have reached (tentative and defeasible as they are) will be less interesting than the method and process of analysis. As noted at the outset, I have used the case study of artificial beings to illustrate how one can conduct a risk asymmetry analysis. In particular, I have emphasised the importance of paying closer attention to the actual evidence regarding the false positive costs and benefits weighed against the false negative costs and benefits. A risk asymmetry analysis should not be simplistic or glib. I believe that this method can be applied to other cases involving relationships with unknown or dubious minds.. I leave it to others to take up the baton and do so.

---

[1] For an overview of the Lemoine/LamDA scandal, see https://www.bbc.com/news/technology-61784011 (accessed March 1, 2023)
[2] Neely, EL. Machines and the Moral Community. *Philosophy and Technology* 2014;27: 97–111.
[3] Sebo, J. The Moral Problem of Other Minds. The Harvard Review of Philosophy 2018; 25:51-70.
[4]Agar, N. How to Treat Machines that Might Have Minds. Philosophy and Technology 2020; 33**:** 269–282 (2020).
[5] Schwitzgebel, E. The Coming Robot Rights Catastrophe, *APA Blog*, 2023 available at https://blog.apaonline.org/2023/01/12/the-coming-robot-rights-catastrophe/ (accessed March 1 2023)
[6] Dennett, D, Kinds of Minds: Towards an Understanding of Consciousness, New York: Basic Books, 1996; and also Hofstadter, D, I am a Strange Loop, New York: Basic Books, 2007.
[7] I have considered these epistemic challenges at greater length in Danaher, J. Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism. Science and Engineering Ethics 2020; 26(4):2023-2049.
[8] For example, see Douglas, KM, Hypersensitive Agency Detection. In: Zeigler-Hill, V., Shackelford, T. eds Encyclopedia of Personality and Individual Differences. Springer, Cham 2017. https://doi.org/10.1007/978-3-319-28099-8_2273-1
[9] For examples, see Grear, A, ed, Should Trees Have Standing? 40 Years On. Cheltenham: Edward Elgar Publishing, 2012.
[10] A small sample of the literature would include: Harris, J & Anthis, JR, The Moral Consideration of Artificial Entities: A Literature Review. Science and Engineering Ethics 2021;27: 53; Véliz, C, Moral zombies: why algorithms are not moral agents. AI and Society, 2021;36: 487–497; Owe, A & Baum, SD, Moral consideration of nonhumans in the ethics of artificial intelligence, AI Ethics 2021; 1:517–528; Mosakas, K, On the moral status of social robots: considering the consciousness criterion. AI and Society 2021; 36: 429–443; Gibert, M & Martin, D, In search of the moral status of AI: why sentience is a strong argument. AI and Society 2022; 37: 319–330; Gordon, J & Gunkel, DJ, Moral Status and Intelligent Robots. Southern Journal of Philosophy 2022; 60: 88–117; Gunkel, D Robot Rights, Cambridge, MA: MIT Press, 2018; Estrada, D, Human supremacy as posthuman risk. *The Journal of Sociotechnical Critique*, 2020; *1*(1): 1-40
[11] Note 10, Gunkel 2018 and also Coeckelbergh, M & Gunkel, DJ, Facing Animals: A Relational, Other-Oriented Approach to Moral Standing. Journal of Agricultural and Environmental Ethics 2014; 27: 715–733.
[12] See note 4 Agar 2019 on this.
[13] See note 2, Neely 2014 -- at least in certain cases. See also note 7, Danaher 2020.
[14] Agar 2019 (note x) suggests that this is only something we can resolve for ourselves, as a practical matter.
[15] The literature on this topic has exploded in recent years. Some key contributions include: Lockhart, T, (2000) Moral Uncertainty and Its Consequences, New York: Oxford, 2000; MacAskill, M, Bykvist, K and Ord, T, Moral uncertainty,, Oxford: Oxford University Press, 2020; Bykvist, K. Moral Uncertainty. Philosophy Compass, 2017; 12(3): e12408; Sepielli, A. Moral uncertainty and the principle of equity among moral theories. Philosophy and Phenomenological Research, 2013; *86*(3): 580-589; Tarsney, C, 2018. Moral

uncertainty for deontologists. Ethical Theory and Moral Practice, 2018; 21: 505-520; Nickel, PJ, Kudina, O and Van de Poel, I, 2022. Moral uncertainty in technomoral change: bridging the explanatory gap. Perspectives on Science, 2022; 30(2): 260-283; Nickel, PJ. Disruptive innovation and moral uncertainty. NanoEthics, 2020;14(3): 259-269; MacAskill, W. Practical ethics given moral uncertainty. Utilitas, 2019; *31*(3): 231-245.

[16] MacAskill, W and Ord, T. Why Maximize Expected Choice-Worthiness? Noûs, 2020; *54*(2): 327-353.

[17] See note 3 Sebo 2018.

[18] Moller, D. Abortion and Moral Risk. Philosophy, 2011; *86*(3): 425-443.

[19] People have made arguments along these lines in relation to artificial minds – see Metzinger, T. Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology. Journal of Artificial Intelligence and Consciousness 2021; 8(1): 43-66.

[20] See note 18, Moller 2011.

[21] Matheson, JD. Applying Moral Caution in the Face of Disagreement. Ethical Theory and Moral Practice, 2021: 1–18 doi:10.1007/s10677-021-10155-x.

[22] See note 4, Agar 2019.

[23] Weatherson, B. Running risks morally. Philosophical Studies 2014: 167 (1):141-163.

[24] See note 21, Matheson 2021.

[25] See note 2, Neely 2014; note 3, Sebo 2018; and note 18, Moller 2011.

[26] Anthis, JR & Paez, E. Moral circle expansion: A promising strategy to impact the far future. Futures 2021; 130: 102756.

[27] Such as: Bryson, JJ. Robots should be slaves. *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues*, 2010: *8*: 63-74; Bryson, JJ. Patiency is not a virtue: the design of intelligent systems and systems of ethics. Ethics and Information Technology, 2018; *20*(1), pp.15-26.

[28] For an overview, see Crimston, CR, Hornsey, MJ, Bain, PG & Bastian, B. Toward a Psychology of Moral Expansiveness. Current Directions in Psychological Science, 2018; 27: 14–19;

[29] Rottman, J, Crimston, CR & Syropoulos, S. Tree-Huggers Versus Human-Lovers: Anthropomorphism and Dehumanization Predict Valuing Nature Over Outgroups. Cognitive Science 2021; 45: e12967.

[30] Such as, note 27 Bryson 2018 and also the critique from Birhane, A and van Dijk, J. 2020, February. Robot rights? Let's talk about human welfare instead. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 2020: 207-213.

[31] See the interview I did with him on this topic, available at https://philosophicaldisquisitions.blogspot.com/2021/11/95-psychology-of-moral-circle.html (accessed March 1 2023)

[32] See note 27, Bryson 2010 and 2018.

[33] See note 5, Schwitzgebel 2023.

[34] Again, it would be impossible to do full justice to the literature on this topic. Representative examples include: Prescott, TJ and Robillard, JM. Are friends electric? The benefits and risks of human-robot relationships. *Iscience*, 2021;*24*(1): 101993; Nyholm, S, Humans and robots: Ethics, agency, and anthropomorphism. Rowman & Littlefield Publishers, 2020; Nyholm, S and Smids, J. Can a robot be a good colleague?. *Science and engineering ethics*, 2020; *26*: 2169-2188; Friedman, C. Ethical concerns with replacing human relations with humanoid robots: an ubuntu perspective. AI and Ethics, 2022: 1-12; Sætra, HS. Loving robots changing love: Towards a practical deficiency-love. Journal of future robot life 2021: (Preprint): pp.1-19; Ryland, H., 2021. It's friendship, Jim, but not as we know it: A degrees-of-friendship view of human–robot friendships. Minds and Machines, 2021;31(3): 377-393; Law, T, Chita-Tegmark, M, Rabb, N and Scheutz, M. Examining attachment to robots: Benefits, challenges, and alternatives. ACM Transactions on Human-Robot Interaction (THRI), 2022; 11(4): 1-18; Cave, S and Dihal, K. AI Will Always Love You: Three Contradictions in Imaginings of Intimate Relations with Machines. In Dainton, Slocombe, Tanyi, eds. Minding the Future: Artificial Intelligence, Philosophical Visions and Science Fiction, Cham: Springer International Publishing, 2021; and Elder, AM, Friendship, robots, and social media: False friends and second selves. London: Routledge, 2017.

[35] Danaher, J, 2019. The philosophical case for robot friendship. *Journal of Posthuman Studies*, 2019:*3*(1): 5-24.

[36] Note 4, Agar 2019 analyses the friendship dilemma in these terms.

[37] See note 34, Ryland 2021.

[38] See references in note 34 above for illustrations of this mindset.

[39] For a discussion see Dunbar, R. How Many Friends Does One Person Need?: Dunbar's Number and Other Evolutionary Quirks, Cambridge, MA: Harvard University Press, 2010; and Dunbar, R, Friends: Understanding the Power of Our Most Important Relationships, London: Little and Brown, 2021.

[40] Lindenfors, P, Wartel, A and Lind, J. 'Dunbar's number'deconstructed. Biology Letters, 2021; *17*(5): 20210158. https://royalsocietypublishing.org/doi/10.1098/rsbl.2021.0158

[41] Dunbar, R Dunbar's number: why my theory that humans can only maintain 150 friendships has withstood 30 years of scrutiny, The Conversation, 2021 - https://theconversation.com/dunbars-number-why-my-theory-that-humans-can-only-maintain-150-friendships-has-withstood-30-years-of-scrutiny-160676 (accessed March 1 2023)
[42] Danaher, J. Robot Betrayal: a guide to the ethics of robotic deception. *Ethics and Information Technology*, 2020;*22*(2); 117-128.
[43] Margalit, A. On betrayal. Cambridge, MA: Harvard University Press, 2017.
[44] See note 7, Danaher 2020.
[45] See note 39, Dunbar 2021, chapter 1.
[46] For example: Leland, J. How loneliness is damaging our health. *New York Times*, 2022, *20*. - https://www.nytimes.com/2022/04/20/nyregion/loneliness-epidemic.html (accessed March 1, 2023)
[47] Murthy, VH, Together: Loneliness, health and what happens when we find connection. New York: Profile Books, 2022.
[48] Dijulio et al 2018, *Loneliness and Social Isolation in the United States, the United Kingdom, and Japan: An International Survey* – available at https://www.kff.org/report-section/loneliness-and-social-isolation-in-the-united-states-the-united-kingdom-and-japan-an-international-survey-introduction/ (accessed March 1, 2023)
[49] Waytz, A, Cacioppo, J, & Epley, N. Who Sees Human? Perspectives in Psychological Science 2010: 5, 219–232; and Powers, KE, Worsham, AL, Freeman, JB, Wheatley, T and Heatherton, TF. Social connection modulates perceptions of animacy. Psychological science 2014; *25*(10): 1943-1948.