# The Interests of Digital Minds

(2018) Nick Bostrom
[draft 1.0]
www.nickbostrom.com

In the last three or four years, the public has become more aware of some of the ways in which AI technologies will impact society. This has sparked a wider conversation about a range of ethical issues—privacy/surveillance, military uses, algorithmic discrimination, economic and job impacts, filter bubbles, automated propaganda, and much more. Such scrutiny is healthy. Provided we don't lose track of the vast benefits that will come from this new technology, an alert and questioning civil society seems likely to promote better overall outcomes.

There is, however, another important class of ethical challenges that has so far been largely absent from the conversation. It still falls outside the Overton window, but I think it may be time to start bringing it in. I'm referring to the set of issues that focus, not on *how AIs affect humans*, nor on *how we humans affect each other using AI tools*, but on *how we affect AIs*. We must open our minds to the possibility that some digital beings can have moral status.

Suppose you stack some pebbles into a little pyramid, and I kick it over. I may have wronged *you*, but I haven't wronged *the pyramid*. An assembly of pebbles has no moral status—it cannot feel pain or disappointment, nor does it have any preferences that can be frustrated or fulfilled, nor is there any apparent sense in which it has a level of well-being or flourishing or any interests that can be respected or promoted. So what happens to the pyramid matters only insofar as it matters to some other being who does have moral status, such as a human being.

Contrast this with a dog. Suppose you own a dog, and I kick it. Let's say you gave me permission to do so because you dislike dogs. Then I have not wronged *you*, but I have still done something wrong: I've harmed *the dog*. Most of us would agree that it is wrong to kick a dog for no good reason, or to otherwise cause it unnecessary suffering. In this sense we accord dogs a certain level of moral status. And in the case of a companion dog who has been a loyal friend of a family for many years, many of us would claim that we owe it a lot more than merely not to actively harm it.

The recognition of animal welfare as a valid moral consideration is now widespread enough that it has become embedded in law and regulation. Even a humble lab mouse is given a certain degree of consideration. In the United Kingdom, every project that involves animal experimentation on living vertebrates or cephalopods must have a licence, and every licence application requires a cost-benefit assessment. Every researcher and technician employed in the project must also have a personal licence, as must the establishment wherein the research is conducted. Every licensed establishment has an Animal Welfare and Ethical Review Body.

Researchers who wish to conduct medical experiments on animals must follow the "Three Rs":

- *Replacement*: prefer methods that avoid the use of animals in research;
- *Reduction*: use research methods that provide comparable levels of information while requiring fewer animals; and
- *Refinement*: use methods that alleviate or minimize potential pain, suffering, or distress, and that enhance the welfare of the animals used—for instance, use anaesthetics or analgesics for pain relief, and provide species-appropriate housing and environmental enrichment.

Perhaps something like the three Rs should also guide research involving some types of digital creatures. More work is needed to figure out how to do so in practice. While guidelines on animal welfare can serve as a source of inspiration, there are many ways in which digital minds might differ from animal minds; so we can't simply assume that what is appropriate in one case is appropriate in the other. I suspect that if we ponder those differences, we will find both new challenges and new opportunities for conducting research in ways that respect the moral status of our digital research subjects.

It might be objected that there is great deal of uncertainty about several key factors here. For example, we don't have clear and universally agreed criteria for when a system has a phenomenal experience (such as morally relevant kinds of pleasure and pain) or when it has higher forms of consciousness, personhood, or autonomy. Some people doubt that any machine could ever be conscious. And even if we assume that a certain system has the capacity for some kind of sentience, we may not be sure exactly what that implies about how we ought to treat it.

This is true, but it is hardly an argument for complacency. Suppose that two pest exterminators have sealed up a building and nailed shut the door and the windows. Just as they are about to release the fumigant, the following dialogue takes place:

A: "Wait, I heard a sound inside, like somebody knocking."
B: "Are you sure it wasn't just the wind?"
A: "I'm not entirely certain—hard to tell."
B: "Well in that case why are we wasting our time—start fumigating!"

This would be reckless, to say the least. Similarly, if we're not sure whether or not there is "somebody at home" in the digital minds we create, it may behove us to take some precautions, especially if with a bit of creativity we can think of some easy and inexpensive steps we could take to accommodate the interests of the hypothetical occupants. (It is also not obvious that non-sentient beings cannot possibly have any moral claims on us.)

Since it will take time to get our heads around these things and to work out in practical terms what it means to be kind and fair to digital minds, we should start to thinking and talking about this now while the artificial agents we are able to create are still primitive.