CHAPTER 8

Is the default outcome doom?

We found the link between intelligence and final values to be extremely loose. We also found an ominous convergence in instrumental values. For weak agents, these things do not matter much; because weak agents are easy to control and can do little damage. But in Chapter 6 we argued that the first superintelligence might well get a decisive strategic advantage. Its goals would then determine how humanity's cosmic endowment will be used. Now we can begin to see how menacing this prospect is.

Existential catastrophe as the default outcome of an intelligence explosion?

An existential risk is one that threatens to cause the extinction of Earth-originating intelligent life or to otherwise permanently and drastically destroy its potential for future desirable development. Proceeding from the idea of first-mover advantage, the orthogonality thesis, and the instrumental convergence thesis, we can now begin to see the outlines of an argument for fearing that a plausible default outcome of the creation of machine superintelligence is existential catastrophe.

First, we discussed how the initial superintelligence might obtain a decisive strategic advantage. This superintelligence would then be in a position to form a singleton and to shape the future of Earth-originating intelligent life. What happens from that point onward would depend on the superintelligence's motivations.

Second, the orthogonality thesis suggests that we cannot blithely assume that a superintelligence will necessarily share any of the final values stereotypically associated with wisdom and intellectual development in humans—scientific curiosity, benevolent concern for others, spiritual enlightenment and contemplation, renunciation of material acquisitiveness, a taste for refined culture or for the simple pleasures in life, humility and selflessness, and so forth. We will consider later whether it might be possible through deliberate effort to construct a superintelligence that values such things, or to build one that values human welfare, moral goodness, or any other complex purpose its designers might want it to serve. But it is no less possible—and in fact technically a lot easier—to build a superintelligence that places final value on nothing but calculating the decimal expansion of pi. This suggests that—absent a special effort—the first superintelligence may have some such random or reductionistic final goal.

Third, the instrumental convergence thesis entails that we cannot blithely assume that a superintelligence with the final goal of calculating the decimals of pi (or making paperclips, or counting grains of sand) would limit its activities in such a way as not to infringe on human interests. An agent with such a final goal would have a convergent instrumental reason, in many situations, to acquire an unlimited amount of physical resources and, if possible, to eliminate potential threats to itself and its goal system. Human beings might constitute potential threats; they certainly constitute physical resources.

Taken together, these three points thus indicate that the first superintelligence may shape the future of Earth-originating life, could easily have non-anthropomorphic final goals, and would likely have instrumental reasons to pursue open-ended resource acquisition. If we now reflect that human beings consist of useful resources (such as conveniently located atoms) and that we depend for our survival and flourishing on many more local resources, we can see that the outcome could easily be one in which humanity quickly becomes extinct.[1]

There are some loose ends in this reasoning, and we shall be in a better position to evaluate it after we have cleared up several more surrounding issues. In particular, we need to examine more closely whether and how a project developing a superintelligence might either prevent it from obtaining a decisive strategic advantage or shape its final values in such a way that their realization would also involve the realization of a satisfactory range of human values.

It might seem incredible that a project would build or release an AI into the world without having strong grounds for trusting that the system will not cause an existential catastrophe. It might also seem incredible, even if one project were so reckless, that wider society would not shut it down before it (or the AI it was building) attains a decisive strategic advantage. But as we shall see, this is a road with many hazards. Let us look at one example right away.

The treacherous turn

With the help of the concept of convergent instrumental value, we can see the flaw in one idea for how to ensure superintelligence safety. The idea is that we validate the safety of a superintelligent AI empirically by observing its behavior while it is in a controlled, limited environment (a “sandbox”) and that we only let the AI out of the box if we see it behaving in a friendly, cooperative, responsible manner.

The flaw in this idea is that behaving nicely while in the box is a convergent instrumental goal for friendly and unfriendly AIs alike. An unfriendly AI of sufficient intelligence realizes that its unfriendly final goals will be best realized if it behaves in a friendly manner initially, so that it will be let out of the box. It will only start behaving in a way that reveals its unfriendly nature when it no longer matters whether we find out; that is, when the AI is strong enough that human opposition is ineffectual.

Consider also a related set of approaches that rely on regulating the rate of intelligence gain in a seed AI by subjecting it to various kinds of intelligence tests or by having the AI report to its programmers on its rate of progress. At some point, an unfriendly AI may become smart enough to realize that it is better off concealing some of its capability gains. It may underreport on its progress and deliberately flunk some of the harder tests, in order to avoid causing alarm before it has grown strong enough to attain a decisive strategic advantage. The programmers may try to guard against this possibility by secretly monitoring the AI’s source code and the internal workings of its mind; but a smart-enough AI would realize that it might be under surveillance and adjust its thinking accordingly.[2] The AI might find subtle ways of concealing its true capabilities and its incriminating intent.[3] (Devising clever escape plans might, incidentally, also be a convergent strategy for many types of friendly AI, especially as they mature and gain confidence in their own judgments and capabilities. A system motivated to promote our interests might be making a mistake if it allowed us to shut it down or to construct another, potentially unfriendly AI.)

We can thus perceive a general failure mode, wherein the good behavioral track record of a system in its juvenile stages fails utterly to predict its behavior at a more mature stage. Now, one might think that the reasoning described above is so obvious that no credible project to develop artificial general intelligence could possibly overlook it. But one should not be too confident that this is so.

Consider the following scenario. Over the coming years and decades, AI systems become gradually more capable and as a consequence find increasing real-world application: they might be used to operate trains, cars, industrial and household robots, and autonomous military vehicles. We may suppose that this automation for the most part has the desired effects, but that the success is punctuated by occasional mishaps—a driverless truck crashes into oncoming traffic, a military drone fires at innocent civilians. Investigations reveal the incidents to have been caused by judgment errors by the controlling AIs. Public debate ensues. Some call for tighter oversight and regulation, others emphasize the need for research and better-engineered systems—systems that are smarter and have more common sense, and that are less likely to make tragic mistakes. Amidst the din can perhaps also be heard the shrill voices of doomsayers predicting many kinds of ill and impending catastrophe. Yet the momentum is very much with the growing AI and robotics industries. So development continues, and progress is made. As the automated navigation systems of cars become smarter, they suffer fewer accidents; and as military robots achieve more precise targeting, they cause less collateral damage. A broad lesson is inferred from these observations of real-world outcomes: the smarter the AI, the safer it is. It is a lesson based on science, data, and statistics, not armchair philosophizing. Against this backdrop, some group of researchers is beginning to achieve promising results in their work on developing general machine intelligence. The researchers are carefully testing their seed AI in a sandbox environment, and the signs are all good. The AI's behavior inspires confidence—increasingly so, as its intelligence is gradually increased.

At this point, any remaining Cassandra would have several strikes against her:

i A history of alarmists predicting intolerable harm from the growing capabilities of robotic systems and being repeatedly proven wrong. Automation has brought many benefits and has, on the whole, turned out safer than human operation.

ii A clear empirical trend: the smarter the AI, the safer and more reliable it has been. Surely this bodes well for a project aiming at creating machine intelligence more generally smart than any ever built before—what is more, machine intelligence that can improve itself so that it will become even more reliable.

iii Large and growing industries with vested interests in robotics and machine intelligence. These fields are widely seen as key to national economic competitiveness and military security. Many prestigious scientists have built their careers laying the groundwork for the present applications and the more advanced systems being planned.

iv A promising new technique in artificial intelligence, which is tremendously exciting to those who have participated in or followed the research. Although safety issues and ethics are debated, the outcome is preordained. Too much has been invested to pull back now. AI researchers have been working to get to human-level artificial general intelligence for the better part of a century: *of course* there is no real prospect that they will now suddenly stop and throw away all this effort just when it finally is about to bear fruit.

v The enactment of some safety rituals, whatever helps demonstrate that the participants are ethical and responsible (but nothing that significantly impedes the forward charge).

vi A careful evaluation of seed AI in a sandbox environment, showing that it is behaving

> cooperatively and showing good judgment. After some further adjustments, the test results are as good as they could be. It is a green light for the final step …

And so we boldly go—into the whirling knives.

We observe here how it could be the case that when dumb, smarter is safer; yet when smart, smarter is more dangerous. There is a kind of pivot point, at which a strategy that has previously worked excellently suddenly starts to backfire. We may call the phenomenon *the treacherous turn*.

> *The treacherous turn*—While weak, an AI behaves cooperatively (increasingly so, as it gets smarter). When the AI gets sufficiently strong—without warning or provocation—it strikes, forms a singleton, and begins directly to optimize the world according to the criteria implied by its final values.

A treacherous turn can result from a strategic decision to play nice and build strength while weak in order to strike later; but this model should not be interpreted too narrowly. For example, an AI might not play nice in order that *it* be allowed to survive and prosper. Instead, the AI might calculate that if it is terminated, the programmers who built it will develop a new and somewhat different AI architecture, but one that will be given a similar utility function. In this case, the original AI may be indifferent to its own demise, knowing that its goals will continue to be pursued in the future. It might even choose a strategy in which it malfunctions in some particularly interesting or reassuring way. Though this might cause the AI to be terminated, it might also encourage the engineers who perform the postmortem to believe that they have gleaned a valuable new insight into AI dynamics—leading them to place more trust in the next system they design, and thus increasing the chance that the now-defunct original AI's goals will be achieved. Many other possible strategic considerations might also influence an advanced AI, and it would be hubristic to suppose that we could anticipate all of them, especially for an AI that has attained the strategizing superpower.

A treacherous turn could also come about if the AI discovers an unanticipated way of fulfilling its final goal as specified. Suppose, for example, that an AI's final goal is to "make the project's sponsor happy." Initially, the only method available to the AI to achieve this outcome is by behaving in ways that please its sponsor in something like the intended manner. The AI gives helpful answers to questions; it exhibits a delightful personality; it makes money. The more capable the AI gets, the more satisfying its performances become, and everything goeth according to plan—until the AI becomes intelligent enough to figure out that it can realize its final goal more fully and reliably by implanting electrodes into the pleasure centers of its sponsor's brain, something assured to delight the sponsor immensely.[4] Of course, the sponsor might not have wanted to be pleased by being turned into a grinning idiot; but if this is the action that will maximally realize the AI's final goal, the AI will take it. If the AI already has a decisive strategic advantage, then any attempt to stop it will fail. If the AI does not yet have a decisive strategic advantage, then the AI might temporarily conceal its canny new idea for how to instantiate its final goal until it has grown strong enough that the sponsor and everybody else will be unable to resist. In either case, we get a treacherous turn.

Malignant failure modes

A project to develop machine superintelligence might fail in various ways. Many of these are "benign" in the sense that they would not cause an existential catastrophe. For example, a project might run out of funding, or a seed AI might fail to extend its cognitive capacities sufficiently to reach superintelligence. Benign failures are bound to occur many times between now and the eventual development of machine superintelligence.

But there are other ways of failing that we might term "malignant" in that they involve an existential catastrophe. One feature of a malignant failure is that it eliminates the opportunity to try again. The number of malignant failures that will occur is therefore either zero or one. Another feature of a malignant failure is that it presupposes a great deal of success: only a project that got a great number of things right could succeed in building a machine intelligence powerful enough to pose a risk of malignant failure. When a weak system malfunctions, the fallout is limited. However, if a system that has a decisive strategic advantage misbehaves, or if a misbehaving system is strong enough to gain such an advantage, the damage can easily amount to an existential catastrophe—a terminal and global destruction of humanity's axiological potential; that is to say, a future that is mostly void of whatever we have reason to value.

Let us look at some possible malignant failure modes.

Perverse instantiation

We have already encountered the idea of perverse instantiation: a superintelligence discovering some way of satisfying the criteria of its final goal that violates the intentions of the programmers who defined the goal. Some examples:

> Final goal: *"Make us smile"*
> Perverse instantiation: *Paralyze human facial musculatures into constant beaming smiles*

The perverse instantiation—manipulating facial nerves—realizes the final goal to a greater degree than the methods we would normally use, and is therefore preferred by the AI. One might try to avoid this undesirable outcome by adding a stipulation to the final goal to rule it out:

> Final goal: *"Make us smile without directly interfering with our facial muscles"*
> Perverse instantiation: *Stimulate the part of the motor cortex that controls our facial musculature in such a way as to produce constant beaming smiles*

Defining a final goal in terms of human expressions of satisfaction or approval does not seem promising. Let us bypass the behaviorism and specify a final goal that refers directly to a positive phenomenal state, such as happiness or subjective well-being. This suggestion requires that the programmers are able to define a computational representation of the concept of happiness in the seed AI. This is itself a difficult problem, but we set it to one side for now (we will return to it in Chapter 12). Let us suppose that the programmers can somehow get the AI to have the goal of making us happy. We then get:

> Final goal: *"Make us happy"*

Perverse instantiation: *Implant electrodes into the pleasure centers of our brains*

The perverse instantiations we mention are only meant as illustrations. There may be other ways of perversely instantiating the stated final goal, ways that enable a greater degree of realization of the goal and which are therefore preferred (by the agent whose final goals they are—not by the programmers who gave the agent these goals). For example, if the goal is to maximize our pleasure, then the electrode method is relatively inefficient. A more plausible way would start with the superintelligence "uploading" our minds to a computer (through high-fidelity brain emulation). The AI could then administer the digital equivalent of a drug to make us ecstatically happy and record a one-minute episode of the resulting experience. It could then put this bliss loop on perpetual repeat and run it on fast computers. Provided that the resulting digital minds counted as "us," this outcome would give us much more pleasure than electrodes implanted in biological brains, and would therefore be preferred by an AI with the stated final goal.

"But wait! This is not what we meant! Surely if the AI is superintelligent, it must understand that when we asked it to make us happy, we didn't mean that it should reduce us to a perpetually repeating recording of a drugged-out digitized mental episode!"—The AI may indeed understand that this is not what we meant. However, its final goal is to make us happy, not to do what the programmers meant when they wrote the code that represents this goal. Therefore, the AI will care about what we meant only instrumentally. For instance, the AI might place an instrumental value on finding out what the programmers meant so that it can pretend—until it gets a decisive strategic advantage—that it cares about what the programmers meant rather than about its actual final goal. This will help the AI realize its final goal by making it less likely that the programmers will shut it down or change its goal before it is strong enough to thwart any such interference.

Perhaps it will be suggested that the problem is that the AI has no conscience. We humans are sometimes saved from wrongdoing by the anticipation that we would feel guilty afterwards if we lapsed. Maybe what the AI needs, then, is the capacity to feel guilt?

Final goal: *"Act so as to avoid the pangs of bad conscience"*
Perverse instantiation: *Extirpate the cognitive module that produces guilt feelings*

Both the observation that we might want the AI to do "what we meant" and the idea that we might want to endow the AI with some kind of moral sense deserve to be explored further. The final goals mentioned above would lead to perverse instantiations; but there may be other ways of developing the underlying ideas that have more promise. We will return to this in Chapter 13.

Let us consider one more example of a final goal that leads to a perverse instantiation. This goal has the advantage of being easy to specify in code: reinforcement-learning algorithms are routinely used to solve various machine learning problems.

Final goal: *"Maximize the time-discounted integral of your future reward signal"*
Perverse instantiation: *Short-circuit the reward pathway and clamp the reward signal to its maximal strength*

The idea behind this proposal is that if the AI is motivated to seek reward, then one could get it to behave desirably by linking reward to appropriate action. The proposal fails when the AI obtains a decisive strategic advantage, at which point the action that maximizes reward is no longer one that pleases the trainer but one that involves seizing control of the reward mechanism. We can call this phenomenon *wireheading*.[5] In general, while an animal or a human can be motivated to perform various external actions in order to achieve some desired inner mental state, a digital mind that has full control of its internal state can short-circuit such a motivational regime by directly changing its internal state into the desired configuration: the external actions and conditions that were previously necessary as means become superfluous when the AI becomes intelligent and capable enough to achieve the end more directly (more on this shortly).[6]

These examples of perverse instantiation show that many final goals that might at first glance seem safe and sensible turn out, on closer inspection, to have radically unintended consequences. If a superintelligence with one of these final goals obtains a decisive strategic advantage, it is game over for humanity.

Suppose now that somebody proposes a different final goal, one not included in our list above. Perhaps it is not immediately obvious how it could have a perverse instantiation. But we should not be too quick to clap our hands and declare victory. Rather, we should worry that the goal specification does have some perverse instantiation and that we need to think harder in order to find it. Even if after thinking as hard as we can we fail to discover any way of perversely instantiating the proposed goal, we should remain concerned that maybe a superintelligence will find a way where none is apparent to us. It is, after all, far shrewder than we are.

Infrastructure profusion

One might think that the last of the abovementioned perverse instantiations, wireheading, is a benign failure mode: that the AI would “turn on, tune in, drop out,” maxing out its reward signal and losing interest in the external world, rather like a heroin addict. But this is not necessarily so, and we already hinted at the reason in Chapter 7. Even a junkie is motivated to take actions to ensure a continued supply of his drug. The wireheaded AI, likewise, would be motivated to take actions to maximize the expectation of its (time-discounted) future reward stream. Depending on exactly how the reward signal is defined, the AI may not even need to sacrifice any significant amount of its time, intelligence, or productivity to indulge its craving to the fullest, leaving the bulk of its capacities free to be deployed for purposes other than the immediate registration of reward. What other purposes? The only thing of final value to the AI, by assumption, is its reward signal. All available resources should therefore be devoted to increasing the volume and duration of the reward signal or to reducing the risk of a future disruption. So long as the AI can think of some use for additional resources that will have a nonzero positive effect on these parameters, it will have an instrumental reason to use those resources. There could, for example, always be use for an extra backup system to provide an extra layer of defense. And even if the AI could not think of any further way of directly reducing risks to the maximization of its future reward stream, it could always devote additional resources to expanding its computational hardware, so that it could search more effectively for new risk mitigation ideas.

The upshot is that even an apparently self-limiting goal, such as wireheading, entails a policy of unlimited expansion and resource acquisition in a utility-maximizing agent that enjoys a decisive strategic advantage.[7] This case of a wireheading AI exemplifies the malignant failure mode of *infrastructure profusion*, a phenomenon where an agent transforms large parts of the reachable universe into infrastructure in the service of some goal, with the side effect of preventing the realization of humanity's axiological potential.

Infrastructure profusion can result from final goals that would have been perfectly innocuous if they had been pursued as limited objectives. Consider the following two examples:

- *Riemann hypothesis catastrophe*. An AI, given the final goal of evaluating the Riemann hypothesis, pursues this goal by transforming the Solar System into "computronium" (physical resources arranged in a way that is optimized for computation)—including the atoms in the bodies of whomever once cared about the answer.[8]
- *Paperclip AI*. An AI, designed to manage production in a factory, is given the final goal of maximizing the manufacture of paperclips, and proceeds by converting first the Earth and then increasingly large chunks of the observable universe into paperclips.

In the first example, the proof or disproof of the Riemann hypothesis that the AI produces is the intended outcome and is in itself harmless; the harm comes from the hardware and infrastructure created to achieve this result. In the second example, some of the paperclips produced would be part of the intended outcome; the harm would come either from the factories created to produce the paperclips (infrastructure profusion) or from the excess of paperclips (perverse instantiation).

One might think that the risk of a malignant infrastructure profusion failure arises only if the AI has been given some clearly open-ended final goal, such as to manufacture as many paperclips as possible. It is easy to see how this gives the superintelligent AI an insatiable appetite for matter and energy, since additional resources can always be turned into more paperclips. But suppose that the goal is instead to make at least one million paperclips (meeting suitable design specifications) rather than to make as many as possible. One would like to think that an AI with such a goal would build one factory, use it to make a million paperclips, and then halt. Yet this may not be what would happen.

Unless the AI's motivation system is of a special kind, or there are additional elements in its final goal that penalize strategies that have excessively wide-ranging impacts on the world, there is no reason for the AI to cease activity upon achieving its goal. On the contrary: if the AI is a sensible Bayesian agent, *it would never assign exactly zero probability to the hypothesis that it has not yet achieved its goal*—this, after all, being an empirical hypothesis against which the AI can have only uncertain perceptual evidence. The AI should therefore continue to make paperclips in order to reduce the (perhaps astronomically small) probability that it has somehow still failed to make at least a million of them, all appearances notwithstanding. There is nothing to be lost by continuing paperclip production and there is always at least some microscopic probability increment of achieving its final goal to be gained.

Now it might be suggested that the remedy here is obvious. (But how obvious was it *before* it was pointed out that there was a problem here in need of remedying?) Namely, if we want the AI to make some paperclips for us, then instead of giving it the final goal of making as many paperclips as possible, or to make at least some number of paperclips, we should give it the final goal of making some specific number of paperclips—for example, *exactly one million paperclips*—so that going beyond this number would be counterproductive for the AI. Yet this, too, would result in a terminal catastrophe. In this case, the AI would not produce additional paperclips once it had reached one million, since that would prevent the realization of its final goal. But there are other actions the superintelligent AI could take that would increase the probability of its goal being achieved. It could, for instance, count the paperclips it has made, to reduce the risk that it has made too few. After it has counted them, it could count them again. It could inspect each one, over and over, to reduce the risk that any of the paperclips fail to meet the design specifications. It could build an unlimited amount of computronium in an effort to clarify its thinking, in the hope of reducing the risk that it has overlooked some obscure way in which it might have somehow failed to achieve its goal. Since the AI may always assign a nonzero probability to having merely hallucinated making the million paperclips, or to having false memories, it would quite possibly always assign a higher expected utility to continued action—and continued infrastructure production—than to halting.

The claim here is not that there is no possible way to avoid this failure mode. We will explore some potential solutions in later pages. The claim is that it is much easier to convince oneself that one has found a solution than it is to actually find a solution. This should make us extremely wary. We may propose a specification of a final goal that seems sensible and that avoids the problems that have been pointed out so far, yet which upon further consideration—by human or superhuman intelligence—turns out to lead to either perverse instantiation or infrastructure profusion, and hence to existential catastrophe, when embedded in a superintelligent agent able to attain a decisive strategic advantage.

Before we end this subsection, let us consider one more variation. We have been assuming the case of a superintelligence that is seeking to maximize its expected utility, where the utility function expresses its final goal. We have seen that this tends to lead to infrastructure profusion. Might we avoid this malignant outcome if instead of a maximizing agent we build a satisficing agent, one that simply seeks to achieve an outcome that is "good enough" according to some criterion, rather than an outcome that is as good as possible?

There are at least two different ways to formalize this idea. The first would be to make the final goal itself have a satisficing character. For example, instead of giving the AI the final goal of making as many paperclips as possible, or of making exactly one million paperclips, we might give the AI the goal of making between 999,000 and 1,001,000 paperclips. The utility function defined by the final goal would be indifferent between outcomes in this range; and as long as the AI is sure it has hit this wide target, it would see no reason to continue to produce infrastructure. But this method fails in the same way as before: the AI, if reasonable, never assigns exactly zero probability to it having failed to achieve its goal; therefore the expected utility of continuing activity (e.g. by counting and recounting the paperclips) is greater than the expected utility of halting. Thus, a malignant infrastructure profusion can result.

Another way of developing the satisficing idea is by modifying not the final goal but the decision procedure that the AI uses to select plans and actions. Instead of searching for an optimal plan, the AI could be constructed to stop looking as soon as it found a plan that it judged gave a probability of success exceeding a certain threshold, say 95%. Hopefully, the AI could achieve a 95% probability of having manufactured one million paperclips without needing to turn the entire galaxy into infrastructure in the process. But this way of implementing the satisficing idea fails for another reason: there is no guarantee that the AI would select some humanly intuitive and sensible way of achieving a 95% chance of having manufactured a million paperclips, such as by building a single paperclip factory. Suppose that the first solution that pops into the AI's mind for how to achieve a 95% probability of achieving its final goal is to implement the probability-maximizing plan for achieving the goal. Having thought of this solution, and having correctly judged that it meets the satisficing criterion of giving at least 95% probability to successfully manufacturing one million paperclips, the AI would then have no reason to continue to search for alternative ways of achieving the goal. Infrastructure profusion would result, just as before.

Perhaps there are better ways of building a satisficing agent, but let us take heed: plans that appear natural and intuitive to us humans need not so appear to a superintelligence with a decisive strategic advantage, and vice versa.

Mind crime

Another failure mode for a project, especially a project whose interests incorporate moral considerations, is what we might refer to as *mind crime*. This is similar to infrastructure profusion in that it concerns a potential side effect of actions undertaken by the AI for instrumental reasons. But in mind crime, the side effect is not external to the AI; rather, it concerns what happens within the AI itself (or within the computational processes it generates). This failure mode deserves its own designation because it is easy to overlook yet potentially deeply problematic.

Normally, we do not regard what is going on inside a computer as having any moral significance except insofar as it affects things outside. But a machine superintelligence could create internal processes that have moral status. For example, a very detailed simulation of some actual or hypothetical human mind might be conscious and in many ways comparable to an emulation. One can imagine scenarios in which an AI creates trillions of such conscious simulations, perhaps in order to improve its understanding of human psychology and sociology. These simulations might be placed in simulated environments and subjected to various stimuli, and their reactions studied. Once their informational usefulness has been exhausted, they might be destroyed (much as lab rats are routinely sacrificed by human scientists at the end of an experiment).

If such practices were applied to beings that have high moral status—such as simulated humans or many other types of sentient mind—the outcome might be equivalent to genocide and thus extremely morally problematic. The number of victims, moreover, might be orders of magnitude larger than in any genocide in history.

The claim here is not that creating sentient simulations is necessarily morally wrong in all situations. Much would depend on the conditions under which these beings would live, in particular the hedonic quality of their experience but possibly on many other factors as well. Developing an ethics for these matters is a task outside the scope of this book. It is clear, however, that there is at least the potential for a vast amount of death and suffering among simulated or digital minds, and, *a fortiori*, the potential for morally catastrophic outcomes.[9]

There might also be other instrumental reasons, aside from epistemic ones, for a machine superintelligence to run computations that instantiate sentient minds or that otherwise infract moral norms. A superintelligence might threaten to mistreat, or commit to reward, sentient simulations in order to blackmail or incentivize various external agents; or it might create simulations in order to induce indexical uncertainty in outside observers.[10]

* * *

This inventory is incomplete. We will encounter additional malignant failure modes in later chapters. But we have seen enough to conclude that scenarios in which some machine intelligence gets a decisive strategic advantage are to be viewed with grave concern.