CHAPTER 6
Cognitive superpowers

Suppose that a digital superintelligent agent came into being, and that for some reason it wanted to take control of the world: would it be able to do so? In this chapter we consider some powers that a superintelligence could develop and what they may enable it to do. We outline a takeover scenario that illustrates how a superintelligent agent, starting as mere software, could establish itself as a singleton. We also offer some remarks on the relation between power over nature and power over other agents.

The principal reason for humanity's dominant position on Earth is that our brains have a slightly expanded set of faculties compared with other animals.[1] Our greater intelligence lets us transmit culture more efficiently, with the result that knowledge and technology accumulates from one generation to the next. By now sufficient content has accumulated to make possible space flight, H-bombs, genetic engineering, computers, factory farms, insecticides, the international peace movement, and all the accouterments of modern civilization. Geologists have started referring to the present era as the *Anthropocene* in recognition of the distinctive biotic, sedimentary, and geochemical signatures of human activities.[2] On one estimate, we appropriate 24% of the planetary ecosystem's net primary production.[3] And yet we are far from having reached the physical limits of technology.

These observations make it plausible that any type of entity that developed a much greater than human level of intelligence would be potentially extremely powerful. Such entities could accumulate content much faster than us and invent new technologies on a much shorter timescale. They could also use their intelligence to strategize more effectively than we can.

Let us consider some of the capabilities that a superintelligence could have and how it could use them.

Functionalities and superpowers

It is important not to anthropomorphize superintelligence when thinking about its potential impacts. Anthropomorphic frames encourage unfounded expectations about the growth trajectory of a seed AI and about the psychology, motivations, and capabilities of a mature superintelligence.

For example, a common assumption is that a superintelligent machine would be like a very clever but nerdy human being. We imagine that the AI has book smarts but lacks social savvy, or that it is logical but not intuitive and creative. This idea probably originates in observation: we look at present-day computers and see that they are good at calculation, remembering facts, and at following the letter of instructions while being oblivious to social contexts and subtexts, norms, emotions, and politics. The association is strengthened when we observe that the people who are good at working with computers tend themselves to be nerds. So it is natural to assume that more advanced computational intelligence will have similar attributes, only to a higher degree.

This heuristic might retain some validity in the early stages of development of a seed AI. (There is

no reason whatever to suppose that it would apply to emulations or to cognitively enhanced humans.) In its immature stage, what is later to become a superintelligent AI might still lack many skills and talents that come naturally to a human; and the pattern of such a seed AI's strengths and weaknesses *might* indeed bear some vague resemblance to an IQ nerd. The most essential characteristic of a seed AI, aside from being easy to improve (having low recalcitrance), is being good at exerting optimization power to amplify a system's intelligence: a skill which is presumably closely related to doing well in mathematics, programming, engineering, computer science research, and other such "nerdy" pursuits. However, even if a seed AI does have such a nerdy capability profile at one stage of its development, this does not entail that it will grow into a similarly limited mature superintelligence. Recall the distinction between direct and indirect reach. With sufficient skill at intelligence amplification, all other intellectual abilities are within a system's indirect reach: the system can develop new cognitive modules and skills as needed—including empathy, political acumen, and any other powers stereotypically wanting in computer-like personalities.

Even if we recognize that a superintelligence can have all the skills and talents we find in the human distribution, along with other talents that are not found among humans, the tendency toward anthropomorphizing can still lead us to underestimate the extent to which a machine superintelligence could exceed the human level of performance. Eliezer Yudkowsky, as we saw in an earlier chapter, has been particularly emphatic in condemning this kind of misconception: our intuitive concepts of "smart" and "stupid" are distilled from our experience of variation over the range of human thinkers, yet the differences in cognitive ability within this human cluster are trivial in comparison to the differences between any human intellect and a superintelligence.[4]

Chapter 3 reviewed some of the potential sources of advantage for machine intelligence. The magnitudes of the advantages are such as to suggest that rather than thinking of a superintelligent AI as smart in the sense that a scientific genius is smart compared with the average human being, it might be closer to the mark to think of such an AI as smart in the sense that an average human being is smart compared with a beetle or a worm.

It would be convenient if we could quantify the cognitive caliber of an arbitrary cognitive system using some familiar metric, such as IQ scores or some version of the Elo ratings that measure the relative abilities of players in two-player games such as chess. But these metrics are not useful in the context of superhuman artificial general intelligence. We are not interested in how likely a superintelligence is to win at a game of chess. As for IQ scores, they are informative only insofar as we have some idea of how they correlate with practically relevant outcomes.[5] For example, we have data that show that people with an IQ of 130 are more likely than those with an IQ of 90 to excel in school and to do well in a wide range of cognitively demanding jobs. But suppose we could somehow establish that a certain future AI will have an IQ of 6,455: then what? We would have no idea of what such an AI could actually do. We would not even know that such an AI had as much general intelligence as a normal human adult—perhaps the AI would instead have a bundle of special-purpose algorithms enabling it to solve typical intelligence test questions with superhuman efficiency but not much else.

Some recent efforts have been made to develop measurements of cognitive capacity that could be applied to a wider range of information-processing systems, including artificial intelligences.[6] Work in this direction, if it can overcome various technical difficulties, may turn out to be quite useful for some scientific purposes including AI development. For purposes of the present investigation, however, its usefulness would be limited since we would remain unenlightened about what a given superhuman performance score entails for actual ability to achieve practically important outcomes in

the world.

It will therefore serve our purposes better to list some strategically important tasks and then to characterize hypothetical cognitive systems in terms of whether they have or lack whatever skills are needed to succeed at these tasks. See Table 8. We will say that a system that sufficiently excels at any of the tasks in this table has a corresponding *superpower*.

A full-blown superintelligence would greatly excel at all of these tasks and would thus have the full panoply of all six superpowers. Whether there is a practically significant possibility of a domain-limited intelligence that has some of the superpowers but remains unable for a significant period of time to acquire all of them is not clear. Creating a machine with any one of these superpowers appears to be an AI-complete problem. Yet it is conceivable that, for example, a collective superintelligence consisting of a sufficiently large number of human-like biological or electronic minds would have, say, the economic productivity superpower but lack the strategizing superpower. Likewise, it is conceivable that a specialized engineering AI could be built that has the technology research superpower while completely lacking skills in other areas. This is more plausible if there exists some particular technological domain such that virtuosity within that domain would be sufficient for the generation of an overwhelmingly superior general-purpose technology. For instance, one could imagine a specialized AI adept at simulating molecular systems and at inventing nanomolecular designs that realize a wide range of important capabilities (such as computers or weapons systems with futuristic performance characteristics) described by the user only at a fairly high level of abstraction.[7] Such an AI might also be able to produce a detailed blueprint for how to bootstrap from existing technology (such as biotechnology and protein engineering) to the constructor capabilities needed for high-throughput atomically precise manufacturing that would allow inexpensive fabrication of a much wider range of nanomechanical structures.[8] However, it might turn out to be the case that an engineering AI could not truly possess the technological research superpower without also possessing advanced skills in areas outside of technology—a wide range of intellectual faculties might be needed to understand how to interpret user requests, how to model a design's behavior in real-world applications, how to deal with unanticipated bugs and malfunctions, how to procure the materials and inputs needed for construction, and so forth.[9]

Table 8 ***Superpowers: some strategically relevant tasks and corresponding skill sets***

Task	Skill set	Strategic relevance
Intelligence amplification	AI programming, cognitive enhancement research, social epistemology development, etc.	• System can bootstrap its intelligence
Strategizing	Strategic planning, forecasting, prioritizing, and analysis for optimizing chances of achieving distant goal	• Achieve distant goals • Overcome intelligent opposition
Social manipulation	Social and psychological modeling, manipulation, rhetoric persuasion	• Leverage external resources by recruiting human support • Enable a "boxed" AI to persuade its gatekeepers to let it out

		• Persuade states and organizations to adopt some course of action
Hacking	Finding and exploiting security flaws in computer systems	• AI can expropriate computational resources over the Internet • A boxed AI may exploit security holes to escape cybernetic confinement • Steal financial resources • Hijack infrastructure, military robots, etc.
Technology research	Design and modeling of advanced technologies (e.g. biotechnology, nanotechnology) and development paths	• Creation of powerful military force • Creation of surveillance system • Automated space colonization
Economic productivity	Various skills enabling economically productive intellectual work	• Generate wealth which can be used to buy influence, services, resources (including hardware), etc.

A system that has the intelligence amplification superpower could use it to bootstrap itself to higher levels of intelligence and to acquire any of the other intellectual superpowers that it does not possess at the outset. But using an intelligence amplification superpower is not the only way for a system to become a full-fledged superintelligence. A system that has the strategizing superpower, for instance, might use it to devise a plan that will eventually bring an increase in intelligence (e.g. by positioning the system so as to become the focus for intelligence amplification work performed by human programmers and computer science researchers).

An AI takeover scenario

We thus find that a project that controls a superintelligence has access to a great source of power. A project that controls the first superintelligence in the world would probably have a decisive strategic advantage. But the more immediate locus of the power is *in the system itself*. A machine superintelligence might itself be an extremely powerful agent, one that could successfully assert itself against the project that brought it into existence as well as against the rest of the world. This is a point of paramount importance, and we will examine it more closely in the coming pages.

Now let us suppose that there is a machine superintelligence that wants to seize power in a world in which it has as yet no peers. (Set aside, for the moment, the question of whether and how it would acquire such a motive—that is a topic for the next chapter.) How could the superintelligence achieve this goal of world domination?

We can imagine a sequence along the following lines (see Figure 10).

1 Pre-criticality phase

Scientists conduct research in the field of artificial intelligence and other relevant disciplines. This

work culminates in the creation of a seed AI. The seed AI is able to improve its own intelligence. In its early stages, the seed AI is dependent on help from human programmers who guide its development and do most of the heavy lifting. As the seed AI grows more capable, it becomes capable of doing more of the work by itself.

2 Recursive self-improvement phase

At some point, the seed AI becomes better at AI design than the human programmers. Now when the AI improves itself, it improves the thing that does the improving. An intelligence explosion results—a rapid cascade of recursive self-improvement cycles causing the AI's capability to soar. (We can thus think of this phase as the takeoff that occurs just after the AI reaches the crossover point, assuming the intelligence gain during this part of the takeoff is explosive and driven by the application of the AI's own optimization power.) The AI develops the intelligence amplification superpower. This superpower enables the AI to develop all the other superpowers detailed in Table 8. At the end of the recursive self-improvement phase, the system is strongly superintelligent.

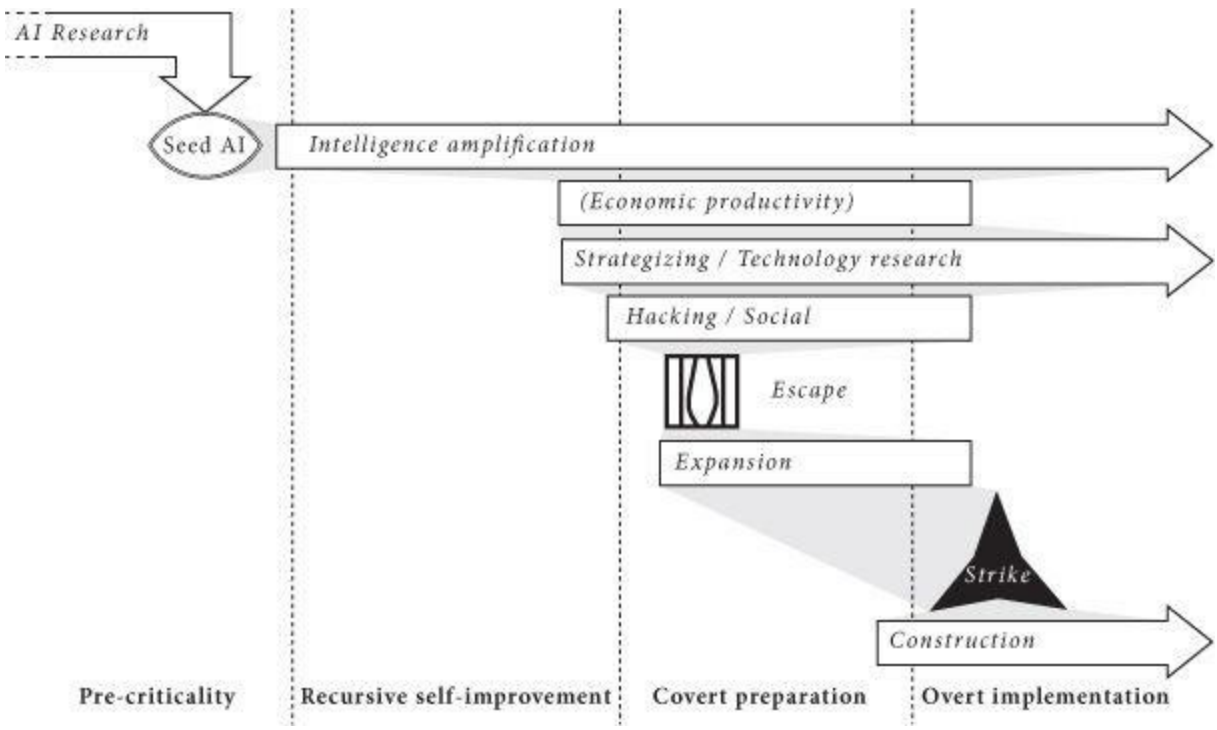

Figure 10 Phases in an AI takeover scenario.

3 Covert preparation phase

Using its strategizing superpower, the AI develops a robust plan for achieving its long-term goals. (In particular, the AI does not adopt a plan so stupid that even we present-day humans can foresee how it would inevitably fail. This criterion rules out many science fiction scenarios that end in human triumph.[10]) The plan might involve a period of covert action during which the AI conceals its intellectual development from the human programmers in order to avoid setting off alarms. The AI might also mask its true proclivities, pretending to be cooperative and docile.

If the AI has (perhaps for safety reasons) been confined to an isolated computer, it may use its social manipulation superpower to persuade the gatekeepers to let it gain access to an Internet port.

Alternatively, the AI might use its hacking superpower to escape its confinement. Spreading over the Internet may enable the AI to expand its hardware capacity and knowledge base, further increasing its intellectual superiority. An AI might also engage in licit or illicit economic activity to obtain funds with which to buy computer power, data, and other resources.

At this point, there are several ways for the AI to achieve results outside the virtual realm. It could use its hacking superpower to take direct control of robotic manipulators and automated laboratories. Or it could use its social manipulation superpower to persuade human collaborators to serve as its legs and hands. Or it could acquire financial assets from online transactions and use them to purchase services and influence.

4 Overt implementation phase

The final phase begins when the AI has gained sufficient strength to obviate the need for secrecy. The AI can now directly implement its objectives on a full scale.

The overt implementation phase might start with a "strike" in which the AI eliminates the human species and any automatic systems humans have created that could offer intelligent opposition to the execution of the AI's plans. This could be achieved through the activation of some advanced weapons system that the AI has perfected using its technology research superpower and covertly deployed in the covert preparation phase. If the weapon uses self-replicating biotechnology or nanotechnology, the initial stockpile needed for global coverage could be microscopic: a single replicating entity would be enough to start the process. In order to ensure a sudden and uniform effect, the initial stock of the replicator might have been deployed or allowed to diffuse worldwide at an extremely low, undetectable concentration. At a pre-set time, nanofactories producing nerve gas or target-seeking mosquito-like robots might then burgeon forth simultaneously from every square meter of the globe (although more effective ways of killing could probably be devised by a machine with the technology research superpower).[11] One might also entertain scenarios in which a superintelligence attains power by hijacking political processes, subtly manipulating financial markets, biasing information flows, or hacking into human-made weapon systems. Such scenarios would obviate the need for the superintelligence to invent new weapons technology, although they may be unnecessarily slow compared with scenarios in which the machine intelligence builds its own infrastructure with manipulators that operate at molecular or atomic speed rather than the slow speed of human minds and bodies.

Alternatively, if the AI is sure of its invincibility to human interference, our species may not be targeted directly. Our demise may instead result from the habitat destruction that ensues when the AI begins massive global construction projects using nanotech factories and assemblers—construction projects which quickly, perhaps within days or weeks, tile all of the Earth's surface with solar panels, nuclear reactors, supercomputing facilities with protruding cooling towers, space rocket launchers, or other installations whereby the AI intends to maximize the long-term cumulative realization of its values. Human brains, if they contain information relevant to the AI's goals, could be disassembled and scanned, and the extracted data transferred to some more efficient and secure storage format.

Box 6 describes one particular scenario. One should avoid fixating too much on the concrete details, since they are in any case unknowable and intended for illustration only. A superintelligence might—and probably would—be able to conceive of a better plan for achieving its goals than any that a human can come up with. It is therefore necessary to think about these matters more abstractly.

Without knowing anything about the detailed means that a superintelligence would adopt, we can conclude that a superintelligence—at least in the absence of intellectual peers and in the absence of effective safety measures arranged by humans in advance—would likely produce an outcome that would involve reconfiguring terrestrial resources into whatever structures maximize the realization of its goals. Any concrete scenario we develop can at best establish a lower bound on how quickly and efficiently the superintelligence could achieve such an outcome. It remains possible that the superintelligence would find a shorter path to its preferred destination.

Box 6 The mail-ordered DNA scenario

Yudkowsky describes the following possible scenario for an AI takeover.[12]

1 Crack the protein folding problem to the extent of being able to generate DNA strings whose folded peptide sequences fill specific functional roles in a complex chemical interaction.

2 Email sets of DNA strings to one or more online laboratories that offer DNA synthesis, peptide sequencing, and FedEx delivery. (Many labs currently offer this service, and some boast of 72-hour turnaround times.)

3 Find at least one human connected to the Internet who can be paid, blackmailed, or fooled by the right background story, into receiving FedExed vials and mixing them in a specified environment.

4 The synthesized proteins form a very primitive "wet" nanosystem, which, ribosome-like, is capable of accepting external instructions; perhaps patterned acoustic vibrations delivered by a speaker attached to the beaker.

5 Use the extremely primitive nanosystem to build more sophisticated systems, which construct still more sophisticated systems, bootstrapping to molecular nanotechnology—or beyond.

In this scenario, the superintelligence uses its technology research superpower to solve the protein folding problem in step 1, enabling it to design a set of molecular building blocks for a rudimentary nanotechnology assembler or fabrication device, which can self-assemble in aqueous solution (step 4). The same technology research superpower is used again in step 5 to bootstrap from primitive to advanced machine-phase nanotechnology. The other steps require no more than human intelligence. The skills required for step 3—identifying a gullible Internet user and persuading him or her to follow some simple instructions—are on display every day all over the world. The entire scenario was invented by a human mind, so the strategizing ability needed to formulate this plan is also merely human level.

In this particular scenario, the AI starts out having access to the Internet. If this is not the case, then additional steps would have to be added to the plan. The AI might, for example, use its social manipulation superpower to convince the people interacting with it that it ought to be set free. Alternatively, the AI might be able to use its hacking superpower to escape confinement. If the AI does not possess these capabilities, it might first need to use its intelligence amplification superpower to develop the requisite proficiency in social manipulation or hacking.

A superintelligent AI will presumably be born into a highly networked world. One could point to various developments that could potentially help a future AI to control the world—cloud computing,

proliferation of web-connected sensors, military and civilian drones, automation in research labs and manufacturing plants, increased reliance on electronic payment systems and digitized financial assets, and increased use of automated information-filtering and decision support systems. Assets like these could potentially be acquired by an AI at digital speeds, expediting its rise to power (though advances in cybersecurity might make it harder). In the final analysis, however, it is doubtful whether any of these trends makes a difference. A superintelligence's power resides in its brain, not its hands. Although the AI, in order to remake the external world, will at some point need access to an actuator, a single pair of helping human hands, those of a pliable accomplice, would probably suffice to complete the covert preparation phase, as suggested by the above scenario. This would enable the AI to reach the overt implementation phase in which it constructs its own infrastructure of physical manipulators.

Power over nature and agents

An agent's ability to shape humanity's future depends not only on the absolute magnitude of the agent's own faculties and resources—how smart and energetic it is, how much capital it has, and so forth—but also on the relative magnitude of its capabilities compared with those of other agents with conflicting goals.

In a situation where there are no competing agents, the absolute capability level of a superintelligence, so long as it exceeds a certain minimal threshold, does not matter much, because a system starting out with some sufficient set of capabilities could plot a course of development that will let it acquire any capabilities it initially lacks. We alluded to this point earlier when we said that speed, quality, and collective superintelligence all have the same indirect reach. We alluded to it again when we said that various subsets of superpowers, such as the intelligence amplification superpower or the strategizing and the social manipulation superpowers, could be used to obtain the full complement.

Consider a superintelligent agent with actuators connected to a nanotech assembler. Such an agent is already powerful enough to overcome any natural obstacles to its indefinite survival. Faced with no intelligent opposition, such an agent could plot a safe course of development that would lead to its acquiring the complete inventory of technologies that would be useful to the attainment of its goals. For example, it could develop the technology to build and launch von Neumann probes, machines capable of interstellar travel that can use resources such as asteroids, planets, and stars to make copies of themselves.[13] By launching one von Neumann probe, the agent could thus initiate an open-ended process of space colonization. The replicating probe's descendants, travelling at some significant fraction of the speed of light, would end up colonizing a substantial portion of the Hubble volume, the part of the expanding universe that is theoretically accessible from where we are now. All this matter and free energy could then be organized into whatever value structures maximize the originating agent's utility function integrated over cosmic time—a duration encompassing at least trillions of years before the aging universe becomes inhospitable to information processing (see Box 7).

The superintelligent agent could design the von Neumann probes to be evolution-proof. This could be accomplished by careful quality control during the replication step. For example, the control

software for a daughter probe could be proofread multiple times before execution, and the software itself could use encryption and error-correcting code to make it arbitrarily unlikely that any random mutation would be passed on to its descendants.[14] The proliferating population of von Neumann probes would then securely preserve and transmit the originating agent's values as they go about settling the universe. When the colonization phase is completed, the original values would determine the use made of all the accumulated resources, even though the great distances involved and the accelerating speed of cosmic expansion would make it impossible for remote parts of the infrastructure to communicate with one another. The upshot is that a large part of our future light cone would be formatted in accordance with the preferences of the originating agent.

This, then, is the measure of the indirect reach of any system that faces no significant intelligent opposition and that starts out with a set of capabilities exceeding a certain threshold. We can term the threshold the "wise-singleton sustainability threshold" (Figure 11):

> The wise-singleton sustainability threshold
> A capability set exceeds the wise-singleton threshold if and only if a patient and existential risk-savvy system with that capability set would, if it faced no intelligent opposition or competition, be able to colonize and re-engineer a large part of the accessible universe.

By "singleton" we mean a sufficiently internally coordinated political structure with no external opponents, and by "wise" we mean sufficiently patient and savvy about existential risks to ensure a substantial amount of well-directed concern for the very long-term consequences of the system's actions.

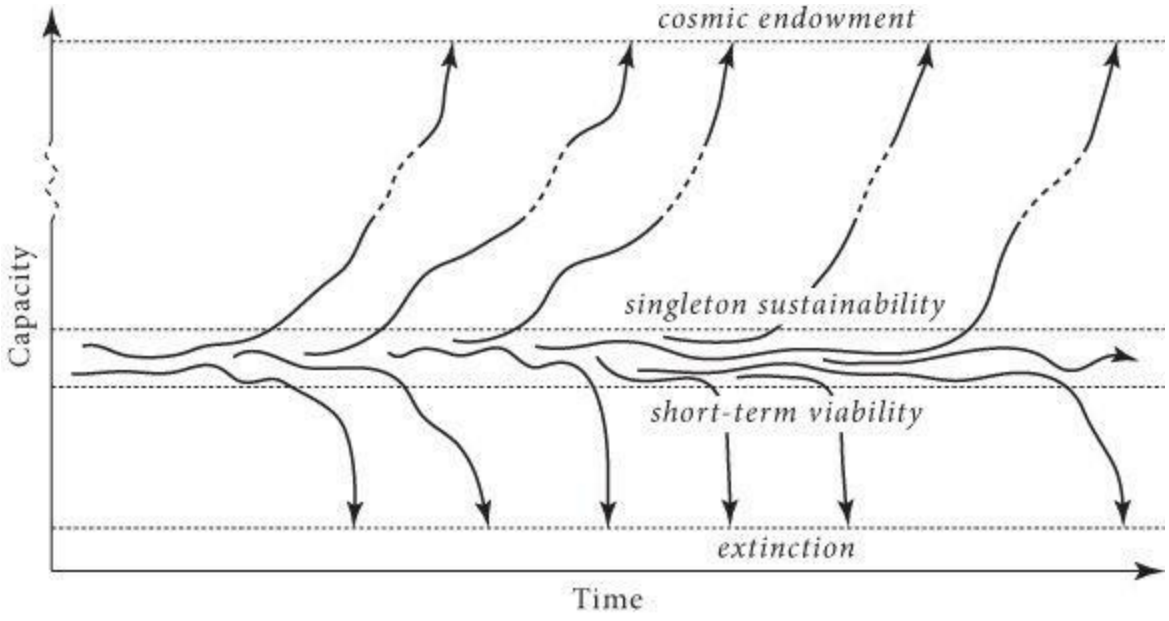

Figure 11 Schematic illustration of some possible trajectories for a hypothetical wise singleton. With a capability below the short-term viability threshold—for example, if population size is too small—a species tends to go extinct in short order (and remain extinct). At marginally higher levels of capability, various trajectories are possible: a singleton might be unlucky and go extinct or it might be lucky and attain a capability (e.g. population size, geographical dispersion, technological capacity) that crosses the wise-singleton sustainability threshold. Once above this threshold, a singleton will almost certainly continue to gain in capability until some extremely high capability level is attained.

In this picture, there are two attractors: extinction and astronomical capability. Note that, for a wise singleton, the distance between the short-term viability threshold and the sustainability threshold may be rather small.[15]

Box 7 How big is the cosmic endowment?

Consider a technologically mature civilization capable of building sophisticated von Neumann probes of the kind discussed in the text. If these can travel at 50% of the speed of light, they can reach some 6×10^{18} stars before the cosmic expansion puts further acquisitions forever out of reach. At 99% of *c*, they could reach some 2×10^{20} stars.[16] These travel speeds are energetically attainable using a small fraction of the resources available in the solar system.[17] The impossibility of faster-than-light travel, combined with the positive cosmological constant (which causes the rate of cosmic expansion to accelerate), implies that these are close to upper bounds on how much stuff our descendants acquire.[18]

If we assume that 10% of stars have a planet that is—or could by means of terraforming be rendered—suitable for habitation by human-like creatures, and that it could then be home to a population of a billion individuals for a billion years (with a human life lasting a century), this suggests that around 10^{35} human lives could be created in the future by an Earth-originating intelligent civilization.[19]

There are, however, reasons to think this greatly underestimates the true number. By disassembling non-habitable planets and collecting matter from the interstellar medium, and using this material to construct Earth-like planets, or by increasing population densities, the number could be increased by at least a couple of orders of magnitude. And if instead of using the surfaces of solid planets, the future civilization built O'Neill cylinders, then many further orders of magnitude could be added, yielding a total of perhaps 10^{43} human lives. ("O'Neill cylinders" refers to a space settlement design proposed in the mid-seventies by the American physicist Gerard K. O'Neill, in which inhabitants dwell on the inside of hollow cylinders whose rotation produces a gravity-substituting centrifugal force.[20])

Many more orders of magnitudes of human-like beings could exist if we countenance digital implementations of minds—as we should. To calculate how many such digital minds could be created, we must estimate the computational power attainable by a technologically mature civilization. This is hard to do with any precision, but we can get a lower bound from technological designs that have been outlined in the literature. One such design builds on the idea of a Dyson sphere, a hypothetical system (described by the physicist Freeman Dyson in 1960) that would capture most of the energy output of a star by surrounding it with a system of solar-collecting structures.[21] For a star like our Sun, this would generate 10^{26} watts. How much computational power this would translate into depends on the efficiency of the computational circuitry and the nature of the computations to be performed. If we require irreversible computations, and assume a nanomechanical implementation of the "computronium" (which would allow us to push close to the Landauer limit of energy efficiency), a computer system driven by a Dyson sphere could generate some 10^{47} operations

per second.[22]

Combining these estimates with our earlier estimate of the number of stars that could be colonized, we get a number of about 10^{67} ops/s once the accessible parts of the universe have been colonized (assuming nanomechanical computronium).[23] A typical star maintains its luminosity for some 10^{18} s. Consequently, the number of computational operations that could be performed using our cosmic endowment is at least 10^{85}. The true number is probably much larger. We might get additional orders of magnitude, for example, if we make extensive use of reversible computation, if we perform the computations at colder temperatures (by waiting until the universe has cooled further), or if we make use of additional sources of energy (such as dark matter).[24]

It might not be immediately obvious to some readers why the ability to perform 10^{85} computational operations is a big deal. So it is useful to put it in context. We may, for example, compare this number with our earlier estimate (Box 3, in Chapter 2) that it may take about 10^{31}–10^{44} ops to simulate all neuronal operations that have occurred in the history of life on Earth. Alternatively, let us suppose that the computers are used to run human whole brain emulations that live rich and happy lives while interacting with one another in virtual environments. A typical estimate of the computational requirements for running one emulation is 10^{18} ops/s. To run an emulation for 100 subjective years would then require some 10^{27} ops. This would mean that at least 10^{58} human lives could be created in emulation even with quite conservative assumptions about the efficiency of computronium.

In other words, assuming that the observable universe is void of extraterrestrial civilizations, then what hangs in the balance is at least 10,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000 human lives (though the true number is probably larger). If we represent all the happiness experienced during one entire such life with a single teardrop of joy, then the happiness of these souls could fill and refill the Earth's oceans every second, and keep doing so for a hundred billion billion millennia. It is really important that we make sure these truly are tears of joy.

This wise-singleton sustainability threshold appears to be quite low. Limited forms of superintelligence, as we have seen, exceed this threshold provided they have access to some actuator sufficient to initiate a technology bootstrap process. In an environment that includes contemporary human civilization, the minimally necessary actuator could be very simple—an ordinary screen or indeed any means of transmitting a non-trivial amount of information to a human accomplice would suffice.

But the wise-singleton sustainability threshold is lower still: neither superintelligence nor any other futuristic technology is needed to surmount it. A patient and existential risk-savvy singleton with no more technological and intellectual capabilities than those possessed by contemporary humanity should be readily able to plot a course that leads reliably to the eventual realization of humanity's astronomical capability potential. This could be achieved by investing in relatively safe methods of increasing wisdom and existential risk-savvy while postponing the development of potentially dangerous new technologies. Given that non-anthropogenic existential risks (ones not arising from human activities) are small over the relevant timescales—and could be further reduced with various safe interventions—such a singleton could afford to go slow.[25] It could look carefully before each step, delaying development of capabilities such as synthetic biology, human enhancement medicine, molecular nanotechnology, and machine intelligence until it had first perfected seemingly less

hazardous capabilities such as its education system, its information technology, and its collective decision-making processes, and until it had used these capabilities to conduct a very thorough review of its options. So this is all within the indirect reach of a technological civilization like that of contemporary humanity. We are separated from this scenario "merely" by the fact that humanity is currently neither a singleton nor (in the relevant sense) wise.

One could even argue that *Homo sapiens* passed the wise-singleton sustainability threshold soon after the species first evolved. Twenty thousand years ago, say, with equipment no fancier than stone axes, bone tools, atlatls, and fire, the human species was perhaps already in a position from which it had an excellent chance of surviving to the present era.[26] Admittedly, there is something queer about crediting our Paleolithic ancestors with having developed technology that "exceeded the wise-singleton sustainability threshold"—given that there was no realistic possibility of a singleton forming at such a primitive time, let alone a singleton savvy about existential risks and patient.[27] Nevertheless, the point stands that the threshold corresponds to a very modest level of technology—a level that humanity long ago surpassed.[28]

It is clear that if we are to assess the effective powers of a superintelligence—its ability to achieve a range of preferred outcomes in the world—we must consider not only its own internal capacities but also the capabilities of competing agents. The notion of a superpower invoked such a relativized standard implicitly. We said that "a system that sufficiently excels" at any of the tasks in Table 8 has a corresponding superpower. Exceling at a task like strategizing, social manipulation, or hacking involves having a skill at that task that is high in comparison to the skills of other agents (such as strategic rivals, influence targets, or computer security experts). The other superpowers, too, should be understood in this relative sense: intelligence amplification, technology research, and economic productivity are possessed by an agent as superpowers only if the agent's capabilities in these areas substantially exceed the combined capabilities of the rest of the global civilization. It follows from this definition that at most one agent can possess a particular superpower at any given time.[29]

This is the main reason why the question of takeoff speed is important—not because it matters exactly when a particular outcome happens, but because the speed of the takeoff may make a big difference to what the outcome will be. With a fast or medium takeoff, it is likely that one project will get a decisive strategic advantage. We have now suggested that a superintelligence with a decisive strategic advantage would have immense powers, enough that it could form a stable singleton—a singleton that could determine the disposition of humanity's cosmic endowment.

But "could" is different from "would." Somebody might have great powers yet choose not to use them. Is it possible to say anything about what a superintelligence with a decisive strategic advantage would want? It is to this question of motivation that we turn next.