CHAPTER 7
The superintelligent will

We have seen that a superintelligence could have a great ability to shape the future according to its goals. But what will its goals be? What is the relation between intelligence and motivation in an artificial agent? Here we develop two theses. The orthogonality thesis holds (with some caveats) that intelligence and final goals are independent variables: any level of intelligence could be combined with any final goal. The instrumental convergence thesis holds that superintelligent agents having any of a wide range of final goals will nevertheless pursue similar intermediary goals because they have common instrumental reasons to do so. Taken together, these theses help us think about what a superintelligent agent would do.

The relation between intelligence and motivation

We have already cautioned against anthropomorphizing the *capabilities* of a superintelligent AI. This warning should be extended to pertain to its *motivations* as well.

It is a useful propaedeutic to this part of our inquiry to first reflect for a moment on the vastness of the space of possible minds. In this abstract space, human minds form a tiny cluster. Consider two persons who seem extremely unlike, perhaps Hannah Arendt and Benny Hill. The personality differences between these two individuals may seem almost maximally large. But this is because our intuitions are calibrated on our experience, which samples from the existing human distribution (and to some extent from fictional personalities constructed by the human imagination for the enjoyment of the human imagination). If we zoom out and consider the space of all possible minds, however, we must conceive of these two personalities as virtual clones. Certainly in terms of neural architecture, Ms. Arendt and Mr. Hill are nearly identical. Imagine their brains lying side by side in quiet repose. You would readily recognize them as two of a kind. You might even be unable to tell which brain belonged to whom. If you looked more closely, studying the morphology of the two brains under a microscope, this impression of fundamental similarity would only be strengthened: you would see the same lamellar organization of the cortex, with the same brain areas, made up of the same types of neuron, soaking in the same bath of neurotransmitters.[1]

Despite the fact that human psychology corresponds to a tiny spot in the space of possible minds, there is a common tendency to project human attributes onto a wide range of alien or artificial cognitive systems. Yudkowsky illustrates this point nicely:

> Back in the era of pulp science fiction, magazine covers occasionally depicted a sentient monstrous alien—colloquially known as a bug-eyed monster (BEM)—carrying off an attractive human female in a torn dress. It would seem the artist believed that a non-humanoid alien, with a wholly different evolutionary history, would sexually desire human females…. Probably the artist did not ask whether a giant bug *perceives* human females as attractive. Rather, a human female in a torn dress *is sexy*—inherently so, as an intrinsic property. They who made this

mistake did not think about the insectoid's mind: they focused on the woman's torn dress. If the dress were not torn, the woman would be less sexy; the BEM does not enter into it.[2]

An artificial intelligence can be far less human-like in its motivations than a green scaly space alien. The extraterrestrial (let us assume) is a biological creature that has arisen through an evolutionary process and can therefore be expected to have the kinds of motivation typical of evolved creatures. It would not be hugely surprising, for example, to find that some random intelligent alien would have motives related to one or more items like food, air, temperature, energy expenditure, occurrence or threat of bodily injury, disease, predation, sex, or progeny. A member of an intelligent social species might also have motivations related to cooperation and competition: like us, it might show in-group loyalty, resentment of free riders, perhaps even a vain concern with reputation and appearance.

Figure 12 Results of anthropomorphizing alien motivation. Least likely hypothesis: space aliens prefer blondes. More likely hypothesis: the illustrators succumbed to the "mind projection fallacy." Most likely hypothesis: the publisher wanted a cover that would entice the target demographic.

An AI, by contrast, need not care intrinsically about any of those things. There is nothing paradoxical about an AI whose sole final goal is to count the grains of sand on Boracay, or to calculate the decimal expansion of pi, or to maximize the total number of paperclips that will exist in its future light cone. In fact, it would be *easier* to create an AI with simple goals like these than to build one that had a human-like set of values and dispositions. Compare how easy it is to write a program that measures how many digits of pi have been calculated and stored in memory with how difficult it would be to create a program that reliably measures the degree of realization of some more meaningful goal—human flourishing, say, or global justice. Unfortunately, because a meaningless reductionistic goal is easier for humans to code and easier for an AI to learn, it is just the kind of goal that a programmer would choose to install in his seed AI if his focus is on taking the quickest path to "getting the AI to work" (without caring much about what exactly the AI will *do*, aside from displaying impressively intelligent behavior). We will revisit this concern shortly.

Intelligent search for instrumentally optimal plans and policies can be performed in the service of any goal. Intelligence and motivation are in a sense orthogonal: we can think of them as two axes spanning a graph in which each point represents a logically possible artificial agent. Some qualifications could be added to this picture. For instance, it might be impossible for a very unintelligent system to have very complex motivations. In order for it to be correct to say that an certain agent "has" a set of motivations, those motivations may need to be functionally integrated with

the agent's decision processes, something that places demands on memory, processing power, and perhaps intelligence. For minds that can modify themselves, there may also be dynamical constraints—an intelligent self-modifying mind with an urgent desire to be stupid might not remain intelligent for long. But these qualifications must not be allowed to obscure the basic point about the independence of intelligence and motivation, which we can express as follows:

> The orthogonality thesis
> Intelligence and final goals are orthogonal: more or less any level of intelligence could in principle be combined with more or less any final goal.

If the orthogonality thesis seems problematic, this might be because of the superficial resemblance it bears to some traditional philosophical positions which have been subject to long debate. Once it is understood to have a different and narrower scope, its credibility should rise. (For example, the orthogonality thesis does not presuppose the Humean theory of motivation.[3] Nor does it presuppose that basic preferences cannot be irrational.[4])

Note that the orthogonality thesis speaks not of *rationality* or *reason*, but of *intelligence*. By "intelligence" we here mean something like skill at prediction, planning, and means–ends reasoning in general.[5] This sense of instrumental cognitive efficaciousness is most relevant when we are seeking to understand what the causal impact of a machine superintelligence might be. Even if there is some (normatively thick) sense of the word "rational" such that a paperclip-maximizing superintelligent agent would necessarily fail to qualify as fully rational in that sense, this would in no way preclude such an agent from having awesome faculties of instrumental reasoning, faculties which could let it have a large impact on the world.[6]

According to the orthogonality thesis, artificial agents can have utterly non-anthropomorphic goals. This, however, does not imply that it is impossible to make predictions about the behavior of particular artificial agents—not even hypothetical superintelligent agents whose cognitive complexity and performance characteristics might render them in some respects opaque to human analysis. There are at least three directions from which we can approach the problem of predicting superintelligent motivation:

- *Predictability through design*. If we can suppose that the designers of a superintelligent agent can successfully engineer the goal system of the agent so that it stably pursues a particular goal set by the programmers, then one prediction we can make is that the agent will pursue that goal. The more intelligent the agent is, the greater the cognitive resourcefulness it will have to pursue that goal. So even before an agent has been created we might be able to predict something about its behavior, if we know something about who will build it and what goals they will want it to have.
- *Predictability through inheritance*. If a digital intelligence is created directly from a human template (as would be the case in a high-fidelity whole brain emulation), then the digital intelligence might inherit the motivations of the human template.[7] The agent might retain some of these motivations even if its cognitive capacities are subsequently enhanced to make it superintelligent. This kind of inference requires caution. The agent's goals and values could easily become corrupted in the uploading process or during its subsequent operation and enhancement, depending on how the procedure is implemented.
- *Predictability through convergent instrumental reasons*. Even without detailed knowledge of an

agent's final goals, we may be able to infer something about its more immediate objectives by considering the *instrumental* reasons that would arise for any of a wide range of possible final goals in a wide range of situations. This way of predicting becomes more useful the greater the intelligence of the agent, because a more intelligent agent is more likely to recognize the true instrumental reasons for its actions, and so act in ways that make it more likely to achieve its goals. (A caveat here is that there might be important instrumental reasons to which *we* are oblivious and which an agent would discover only once it reaches some very high level of intelligence—this could make the behavior of superintelligent agents less predictable.)

The next section explores this third way of predictability and develops an "instrumental convergence thesis" which complements the orthogonality thesis. Against this background we can then better examine the other two sorts of predictability, which we will do in later chapters where we ask what might be done to shape an intelligence explosion to increase the chances of a beneficial outcome.

Instrumental convergence

According to the orthogonality thesis, intelligent agents may have an enormous range of possible final goals. Nevertheless, according to what we may term the "instrumental convergence" thesis, there are some *instrumental* goals likely to be pursued by almost any intelligent agent, because there are some objectives that are useful intermediaries to the achievement of almost any final goal. We can formulate this thesis as follows:

> The instrumental convergence thesis
> Several instrumental values can be identified which are convergent in the sense that their attainment would increase the chances of the agent's goal being realized for a wide range of final goals and a wide range of situations, implying that these instrumental values are likely to be pursued by a broad spectrum of situated intelligent agents.

In the following we will consider several categories where such convergent instrumental values may be found.[8] The likelihood that an agent will recognize the instrumental values it confronts increases (*ceteris paribus*) with the agent's intelligence. We will therefore focus mainly on the case of a hypothetical superintelligent agent whose instrumental reasoning capacities far exceed those of any human. We will also comment on how the instrumental convergence thesis applies to the case of human beings, as this gives us occasion to elaborate some essential qualifications concerning how the instrumental convergence thesis should be interpreted and applied. Where there are convergent instrumental values, we may be able to predict some aspects of a superintelligence's behavior even if we know virtually nothing about that superintelligence's final goals.

Self-preservation

If an agent's final goals concern the future, then in many scenarios there will be future actions it could perform to increase the probability of achieving its goals. This creates an instrumental reason for the

agent to try to be around in the future—to help achieve its future-oriented goal.

Most humans seem to place some *final* value on their own survival. This is not a necessary feature of artificial agents: some may be designed to place no final value whatever on their own survival. Nevertheless, many agents that do not care intrinsically about their own survival would, under a fairly wide range of conditions, care instrumentally about their own survival in order to accomplish their final goals.

Goal-content integrity

If an agent retains its present goals into the future, then its present goals will be more likely to be achieved by its future self. This gives the agent a present instrumental reason to prevent alterations of its final goals. (The argument applies only to final goals. In order to attain its final goals, an intelligent agent will of course routinely want to change its *subgoals* in light of new information and insight.)

Goal-content integrity for final goals is in a sense even more fundamental than survival as a convergent instrumental motivation. Among humans, the opposite may seem to hold, but that is because survival is usually part of our final goals. For software agents, which can easily switch bodies or create exact duplicates of themselves, preservation of self as a particular implementation or a particular physical object need not be an important instrumental value. Advanced software agents might also be able to swap memories, download skills, and radically modify their cognitive architecture and personalities. A population of such agents might operate more like a "functional soup" than a society composed of distinct semi-permanent persons.[9] For some purposes, processes in such a system might be better individuated as *teleological threads*, based on their values, rather than on the basis of bodies, personalities, memories, or abilities. In such scenarios, goal-continuity might be said to *constitute* a key aspect of survival.

Even so, there are situations in which an agent can best fulfill its final goals by intentionally changing them. Such situations can arise when any of the following factors is significant:

- *Social signaling*. When others can perceive an agent's goals and use that information to infer instrumentally relevant dispositions or other correlated attributes, it can be in the agent's interest to modify its goals to make a favorable impression. For example, an agent might miss out on beneficial deals if potential partners cannot trust it to fulfill its side of the bargain. In order to make credible commitments, an agent might therefore wish to adopt as a final goal the honoring of its earlier commitments (and allow others to verify that it has indeed adopted this goal). Agents that could flexibly and transparently modify their own goals could use this ability to enforce deals.[10]
- *Social preferences*. Others may also have final preferences about an agent's goals. The agent could then have reason to modify its goals, either to satisfy or to frustrate those preferences.
- *Preferences concerning own goal content*. An agent might have some final goal concerned with the agent's own goal content. For example, the agent might have a final goal to become the type of agent that is motivated by certain values rather than others (such as compassion rather than comfort).
- *Storage costs*. If the cost of storing or processing some part of an agent's utility function is large compared to the chance that a situation will arise in which applying that part of the utility function will make a difference, then the agent has an instrumental reason to simplify its goal content, and it

may trash the bit that is idle.[11]

We humans often seem happy to let our final values drift. This might often be because we do not know precisely what they are. It is not surprising that we want our *beliefs* about our final values to be able to change in light of continuing self-discovery or changing self-presentation needs. However, there are cases in which we willingly change the values themselves, not just our beliefs or interpretations of them. For example, somebody deciding to have a child might predict that they will come to value the child for its own sake, even though at the time of the decision they may not particularly value their future child or like children in general.

Humans are complicated, and many factors might be at play in a situation like this.[12] For instance, one might have a final value that involves becoming the kind of person who cares about some other individual for his or her own sake, or one might have a final value that involves having certain experiences and occupying a certain social role; and becoming a parent—and undergoing the attendant goal shift—might be a necessary aspect of that. Human goals can also have inconsistent content, and so some people might want to modify some of their final goals to reduce the inconsistencies.

Cognitive enhancement

Improvements in rationality and intelligence will tend to improve an agent's decision-making, rendering the agent more likely to achieve its final goals. One would therefore expect cognitive enhancement to emerge as an instrumental goal for a wide variety of intelligent agents. For similar reasons, agents will tend to instrumentally value many kinds of information.[13]

Not all kinds of rationality, intelligence, and knowledge need be instrumentally useful in the attainment of an agent's final goals. "Dutch book arguments" can be used to show that an agent whose credence function violates the rules of probability theory is susceptible to "money pump" procedures, in which a savvy bookie arranges a set of bets each of which appears favorable according to the agent's beliefs, but which in combination are guaranteed to result in a loss for the agent, and a corresponding gain for the bookie.[14] However, this fact fails to provide any strong general instrumental reasons to iron out all probabilistic incoherency. Agents who do not expect to encounter savvy bookies, or who adopt a general policy against betting, do not necessarily stand to lose much from having some incoherent beliefs—and they may gain important benefits of the types mentioned: reduced cognitive effort, social signaling, etc. There is no general reason to expect an agent to seek instrumentally useless forms of cognitive enhancement, as an agent might not value knowledge and understanding for their own sakes.

Which cognitive abilities are instrumentally useful depends both on the agent's final goals and on its situation. An agent that has access to reliable expert advice may have little need for its own intelligence and knowledge. If intelligence and knowledge come at a cost, such as time and effort expended in acquisition, or increased storage or processing requirements, then the agent might prefer less knowledge and less intelligence.[15] The same can hold if the agent has final goals that involve being ignorant of certain facts; and likewise if an agent faces incentives arising from strategic commitments, signaling, or social preferences.[16]

Each of these countervailing reasons often comes into play for human beings. Much information is

irrelevant to our goals; we can often rely on others' skill and expertise; acquiring knowledge takes time and effort; we might intrinsically value certain kinds of ignorance; and we operate in an environment in which the ability to make strategic commitments, socially signal, and satisfy other people's direct preferences over our own epistemic states is often more important to us than simple cognitive gains.

There are special situations in which cognitive enhancement may result in an enormous increase in an agent's ability to achieve its final goals—in particular, if the agent's final goals are fairly unbounded and the agent is in a position to become the first superintelligence and thereby potentially obtain a decisive strategic advantage, enabling the agent to shape the future of Earth-originating life and accessible cosmic resources according to its preferences. At least in this special case, a rational intelligent agent would place a very high instrumental value on cognitive enhancement.

Technological perfection

An agent may often have instrumental reasons to seek better technology, which at its simplest means seeking more efficient ways of transforming some given set of inputs into valued outputs. Thus, a software agent might place an instrumental value on more efficient algorithms that enable its mental functions to run faster on given hardware. Similarly, agents whose goals require some form of physical construction might instrumentally value improved engineering technology which enables them to create a wider range of structures more quickly and reliably, using fewer or cheaper materials and less energy. Of course, there is a tradeoff: the potential benefits of better technology must be weighed against its costs, including not only the cost of obtaining the technology but also the costs of learning how to use it, integrating it with other technologies already in use, and so forth.

Proponents of some new technology, confident in its superiority to existing alternatives, are often dismayed when other people do not share their enthusiasm. But people's resistance to novel and nominally superior technology need not be based on ignorance or irrationality. A technology's valence or normative character depends not only on the context in which it is deployed, but also the vantage point from which its impacts are evaluated: what is a boon from one person's perspective can be a liability from another's. Thus, although mechanized looms increased the economic efficiency of textile production, the Luddite handloom weavers who anticipated that the innovation would render their artisan skills obsolete may have had good instrumental reasons to oppose it. The point here is that if "technological perfection" is to name a widely convergent instrumental goal for intelligent agents, then the term must be understood in a special sense—technology must be construed as embedded in a particular social context, and its costs and benefits must be evaluated with reference to some specified agents' final values.

It seems that a superintelligent *singleton*—a superintelligent agent that faces no significant intelligent rivals or opposition, and is thus in a position to determine global policy unilaterally—would have instrumental reason to perfect the technologies that would make it better able to shape the world according to its preferred designs.[17] This would probably include space colonization technology, such as von Neumann probes. Molecular nanotechnology, or some alternative still more capable physical manufacturing technology, also seems potentially very useful in the service of an extremely wide range of final goals.[18]

Resource acquisition

Finally, resource acquisition is another common emergent instrumental goal, for much the same reasons as technological perfection: both technology and resources facilitate physical construction projects.

Human beings tend to seek to acquire resources sufficient to meet their basic biological needs. But people usually seek to acquire resources far beyond this minimum level. In doing so, they may be partially driven by lesser physical desiderata, such as increased convenience. A great deal of resource accumulation is motivated by social concerns—gaining status, mates, friends, and influence, through wealth accumulation and conspicuous consumption. Perhaps less commonly, some people seek additional resources to achieve altruistic ambitions or expensive non-social aims.

On the basis of such observations it might be tempting to suppose that a superintelligence not facing a competitive social world would see no instrumental reason to accumulate resources beyond some modest level, for instance whatever computational resources are needed to run its mind along with some virtual reality. Yet such a supposition would be entirely unwarranted. First, the value of resources depends on the uses to which they can be put, which in turn depends on the available technology. With mature technology, basic resources such as time, space, matter, and free energy, could be processed to serve almost any goal. For instance, such basic resources could be converted into life. Increased computational resources could be used to run the superintelligence at a greater speed and for a longer duration, or to create additional physical or simulated lives and civilizations. Extra physical resources could also be used to create backup systems or perimeter defenses, enhancing security. Such projects could easily consume far more than one planet's worth of resources.

Furthermore, the cost of acquiring additional extraterrestrial resources will decline radically as the technology matures. Once von Neumann probes can be built, a large portion of the observable universe (assuming it is uninhabited by intelligent life) could be gradually colonized—for the one-off cost of building and launching a single successful self-reproducing probe. This low cost of celestial resource acquisition would mean that such expansion could be worthwhile even if the value of the additional resources gained were somewhat marginal. For example, even if a superintelligence's final goals only concerned what happened within some particular small volume of space, such as the space occupied by its original home planet, it would still have instrumental reasons to harvest the resources of the cosmos beyond. It could use those surplus resources to build computers to calculate more optimal ways of using resources within the small spatial region of primary concern. It could also use the extra resources to build ever more robust fortifications to safeguard its sanctum. Since the cost of acquiring additional resources would keep declining, this process of optimizing and increasing safeguards might well continue indefinitely even if it were subject to steeply diminishing returns.[19]

Thus, there is an extremely wide range of possible final goals a superintelligent singleton could have that would generate the instrumental goal of unlimited resource acquisition. The likely manifestation of this would be the superintelligence's initiation of a colonization process that would expand in all directions using von Neumann probes. This would result in an approximate sphere of expanding infrastructure centered on the originating planet and growing in radius at some fraction of the speed of light; and the colonization of the universe would continue in this manner until the accelerating speed of cosmic expansion (a consequence of the positive cosmological constant) makes

further procurements impossible as remoter regions drift permanently out of reach (this happens on a timescale of billions of years).[20] By contrast, agents lacking the technology required for inexpensive resource acquisition, or for the conversion of generic physical resources into useful infrastructure, may often find it not cost-effective to invest any present resources in increasing their material endowments. The same may hold for agents operating in competition with other agents of similar powers. For instance, if competing agents have already secured accessible cosmic resources, there may be no colonization opportunities left for a late-starting agent. The convergent instrumental reasons for superintelligences uncertain of the non-existence of other powerful superintelligent agents are complicated by strategic considerations that we do not currently fully understand but which may constitute important qualifications to the examples of convergent instrumental reasons we have looked at here.[21]

* * *

It should be emphasized that the existence of convergent instrumental reasons, even if they apply to and are recognized by a particular agent, does not imply that the agent's behavior is easily predictable. An agent might well think of ways of pursuing the relevant instrumental values that do not readily occur to us. This is especially true for a superintelligence, which could devise extremely clever but counterintuitive plans to realize its goals, possibly even exploiting as-yet undiscovered physical phenomena.[22] What is predictable is that the convergent instrumental values would be pursued and used to realize the agent's final goals—not the specific actions that the agent would take to achieve this.