23

The Control Problem. Excerpts from *Superintelligence: Paths, Dangers, Strategies*

Nick Bostrom

If we are threatened with existential catastrophe as the default outcome of an intelligence explosion, our thinking must immediately turn to the search for countermeasures. Is there some way to avoid the default outcome? Is it possible to engineer a controlled detonation? In this chapter we begin to analyze the control problem, the unique principal-agent problem that arises with the creation of an artificial superintelligent agent. We distinguish two broad classes of potential methods for addressing this problem – capability control and motivation selection – and we examine several specific techniques within each class. We also allude to the esoteric possibility of "anthropic capture."

Two Agency Problems

If we suspect that the default outcome of an intelligence explosion is existential catastrophe, our thinking must immediately turn to whether, and if so how, this default outcome can be avoided. Is it possible to achieve a "controlled detonation"? Could we engineer the initial conditions of an intelligence explosion so as to achieve a specific desired outcome, or at least to ensure that the result lies somewhere in the class of broadly acceptable outcomes? More specifically: how can the sponsor of a project that aims to develop superintelligence ensure that the project, if successful, produces a

Original publication details: "The Control Problem," Nick Bostrom, *Superintelligence*, Oxford University Press, 2014, pp. 127–144. By permission of Oxford University Press.

Science Fiction and Philosophy: From Time Travel to Superintelligence, Second Edition. Edited by Susan Schneider.

superintelligence that would realize the sponsor's goals? We can divide this control problem into two parts. One part is generic, the other unique to the present context.

This first part – what we shall call the *first principal–agent problem* – arises whenever some human entity ("the principal") appoints another ("the agent") to act in the former's interest. This type of agency problem has been extensively studied by economists.[1] It becomes relevant to our present concern if the people creating an AI are distinct from the people commissioning its creation. The project's owner or sponsor (which could be anything ranging from a single individual to humanity as a whole) might then worry that the scientists and programmers implementing the project will not act in the sponsor's best interest.[2] Although this type of agency problem could pose significant challenges to a project sponsor, it is not a problem unique to intelligence amplification or AI projects. Principal–agent problems of this sort are ubiquitous in human economic and political interactions, and there are many ways of dealing with them. For instance, the risk that a disloyal employee will sabotage or subvert the project could be minimized through careful background checks of key personnel, the use of a good version control system for software projects, and intensive oversight from multiple independent monitors and auditors. Of course, such safeguards come at a cost – they expand staffing needs, complicate personnel selection, hinder creativity, and stifle independent and critical thought, all of which could reduce the pace of progress. These costs could be significant, especially for projects that have tight budgets, or that perceive themselves to be in a close race in a winner-takes-all competition. In such situations, projects may skimp on procedural safeguards, creating possibilities for potentially catastrophic principal-agent failures of the first type.

The other part of the control problem is more specific to the context of an intelligence explosion. This is the problem that a project faces when it seeks to ensure that the superintelligence it is building will not harm the project's interests This part, too, can be thought of as a principal–agent problem – the *second principal–agent problem* In this case, the agent is not a human agent operating on behalf of a human principal. Instead, the agent is the superintelligent system. Whereas the first principal–agent problem occurs mainly in the development phase, the second agency problem threatens to cause trouble mainly in the superintelligence's operational phase.

Exhibit I Two agency problems

The first principal–agent problem

- Human v. Human (Sponsor → Developer)
- Occurs mainly in developmental phase
- Standard management techniques apply

The second principal–agent problem ("the control problem")
- Human v. Superintelligence (Project → System)
- Occurs mainly in operational (and bootstrap) phase
- New techniques needed

This second agency problem poses an unprecedented challenge. Solving it will require new techniques. We have already considered some of the difficulties involved. We saw, in particular, that the treacherous turn syndrome vitiates what might otherwise have seemed like a promising set of methods, ones that rely on observing an AI's behavior in its developmental phase and allowing the AI to graduate from a secure environment once it has accumulated a track record of taking appropriate actions. Other technologies can often be safety-tested in the laboratory or in small field studies, and then rolled out gradually with a possibility of halting deployment if unexpected troubles arise. Their performance in preliminary trials helps us make reasonable inferences about their future reliability. Such behavioral methods are defeated in the case of superintelligence because of the strategic planning ability of general intelligence.[3]

Since the behavioral approach is unavailing, we must look for alternatives. We can divide potential control methods into two broad classes: *capability control methods*, which aim to control what the superintelligence can do; and *motivation selection methods*, which aim to control what it wants to do. Some of the methods are compatible while others represent mutually exclusive alternatives. In this chapter we canvass the main options. …

It is important to realize that some control method (or combination of methods) must be implemented *before* the system becomes superintelligent. It cannot be done after the system has obtained a decisive strategic advantage. The need to solve the control problem in advance – and to implement the solution successfully in the very first system to attain superintelligence – is part of what makes achieving a controlled detonation such a daunting challenge.

Capability Control Methods

Capability control methods seek to prevent undesirable outcomes by limiting what the superintelligence can do. This might involve placing the superintelligence in an environment in which it is unable to cause harm (*boxing methods*) or in which there are strongly convergent instrumental reasons not to engage in harmful behavior (*incentive methods*). It might also involve limiting the internal capacities of the superintelligence (*stunting*). In addition, capability control methods might involve the use of mechanisms to automatically detect and react to various kinds of containment failure or attempted transgression (*tripwires*).

Boxing Methods

Boxing methods can be subdivided into physical and informational containment methods.

Physical containment aims to confine the system to a "box," i.e. to prevent the system from interacting with the external world otherwise than via specific restricted output channels. The boxed system would not have access to physical manipulators outside of the box. Removing manipulators (such as robotic arms) from inside the box as well would prevent the system from constructing physical devices that could breach the confinement.

For extra security, the system should be placed in a metal mesh to prevent it from transmitting radio signals, which might otherwise offer a means of manipulating electronic objects such as radio receivers in the environment. Note, by the way, how easy it might have been to overlook the need for this precaution. One might naively have assumed that an agent without a manipulator could not affect the external world. But it might be possible for a machine intelligence to generate radio waves even when it lacks access to external manipulators, simply "by thinking" (that is, by shuffling the electrons in its circuitry in particular patterns).[4] Once pointed out, this vulnerability can be patched by enclosing the system in a Faraday cage – but we are left wondering how many other similarly subtle vulnerabilities might exist. Each time we hear of a seemingly foolproof security design that has an unexpected flaw, we should prick up our ears. These occasions grace us with the opportunity to abandon a life of overconfidence and resolve to become better Bayesians.[5]

Physical confinement has several advantages. It is easy to implement. It can be applied to many machine intelligence architectures, even ones that were not initially designed with safety as an objective. It can be used in combination with most other control methods. And it seems unlikely to go wrong by backfiring: that is, while it might fail to ensure safety, it is unlikely to cause a catastrophe that would not otherwise have occurred. The main disadvantage with physical confinement is that it reduces the functionality of the superintelligence. … Another concern is that it might encourage a false sense of security, though this is avoidable if we regard physical confinement as icing on the cake rather than the main substance of our precautions.

Informational containment aims to restrict what information is allowed to exit the box. We have already seen how a superintelligence that has access to an Internet port, such that it can message outside entities, is potentially unsafe: even if it starts out without access to physical actuators, it may use its information output channel to get human beings to do its bidding. An obvious informational containment method, therefore, is to bar the system from accessing communications networks.

Unplugging an AI from the Internet does not, however, ensure safety if there are one or more humans serving as the system's gatekeepers and

remaining in communication with it. During the development phase, the programmers serve as gatekeepers; during the operational phase, the system's owner might play this role. If the AI can persuade or trick a gatekeeper to let it out of the box, resulting in its gaining access either to the Internet or directly to physical manipulators, then the boxing strategy has failed. Human beings are not secure systems, especially not when pitched against a superintelligent schemer and persuader.[6]

Although routing the outbound communications of a machine intelligence to a human gatekeeper does not guarantee safety, it could serve as an auxiliary safety measure. There are precautions that might reduce the risk of the AI escaping through the gatekeeper. The AI could be allowed to communicate only via a low-bandwidth teletype interface. This would make it slightly harder to psychologically manipulate the gatekeepers (precluding the use of emotive sound and imagery and tricks such as flashing hypnotic patterns to induce a trancelike state).[7] Going further, the content (not just the mode of presentation) of the output could be restricted; so that, for example, the AI would be constrained to giving single-bit answers to yes/no questions. …

The limiting case of the boxing approach would be a system kept in complete physical and informational isolation. Even if achievable, however, such an isolated system would be rather useless since it would have no effect on the external world. It might perhaps be thought that some scientific purpose could be served by creating a superintelligence and keeping it in isolation: by studying a self-contained model system, one could learn about its internal dynamics and its behavior patterns. But this would be an error. As soon as the system is observed, it ceases to be informationally isolated.[8] Information can be transmitted not only via messages that an AI sends out through a designated "output channel" but also via any observation an outsider makes of some causal consequence of the AI's workings, direct or indirect – its power consumption, its CPU and memory usage, its computational states, or indeed any traces left behind after it has been shut down. An AI anticipating that it might be observed in any of these ways could strategically adopt behaviors designed to influence the hypothesized observers. For example, the AI could "malfunction" in whichever way would most increase the probability that it will ultimately be reactivated or that another superintelligence will be created with goals similar to its own.

Incentive Methods

Incentive methods involve placing an agent in an environment where it finds instrumental reasons to act in ways that promote the principal's interests.

Consider a billionaire who uses her fortune to set up a large charitable foundation. Once created, the foundation may be powerful – more powerful

than most individuals, including its founder, who might have donated most of her wealth. To control the foundation, the founder lays down its purpose in articles of incorporation and bylaws, and appoints a board of directors sympathetic to her cause. These measures constitute a form of motivation selection, since they aim to shape foundation's preferences. But even if such attempts to customize the organizational internals fail, the foundation's behavior would remain circumscribed by its social and legal milieu. The foundation would have an incentive to obey the law, for example, lest it be shut down or fined. It would have an incentive to offer its employees acceptable pay and working conditions, and to satisfy external stakeholders. Whatever its final goals, the foundation thus has instrumental reasons to conform its behavior to various social norms.

Might one not hope that a machine superintelligence would likewise be hemmed in by the need to get along with the other actors with which it shares the stage? Though this might seem like a straightforward way of dealing with the control problem, it is not free of obstacles. In particular, it presupposes a balance of power: legal or economic sanctions cannot restrain an agent that has a decisive strategic advantage. Social integration can therefore not be relied upon as a control method in fast or medium takeoff scenarios that feature a winner-takes-all dynamic.

How about in multipolar scenarios, wherein several agencies emerge post-transition with comparable levels of capability? Unless the default trajectory is one with a slow takeoff, achieving such a power distribution may require a carefully orchestrated ascent wherein different projects are deliberately synchronized to prevent any one of them from ever pulling ahead of the pack.[9] Even if a multipolar outcome does result, social integration is not a perfect solution. By relying on social integration to solve the control problem, the principal risks sacrificing a large portion of her potential influence. Although a balance of power might prevent a particular AI from taking over the world, that AI will still have *some* power to affect outcomes; and if that power is used to promote some arbitrary final goal – maximizing paperclip production – it is probably not being used to advance the interests of the principal. Imagine our billionaire endowing a new foundation and allowing its mission to be set by a random word generator: not a species-level threat, but surely a wasted opportunity.

A related but importantly different idea is that an AI, by interacting freely in society, would acquire new human-friendly final goals. Some such process of socialization takes place in us humans. We internalize norms and ideologies, and we come to value other individuals for their own sakes in consequence of our experiences with them. But this is not a universal dynamic present in all intelligent systems. As discussed earlier, many types of agent in many situations will have convergent instrumental reasons *not* to permit changes in their final goals. (One might consider trying to design a

special kind of goal system that can acquire final goals in the manner that humans do; but this would not count as a capability control method. …)

Capability control through social integration and balance of power relies upon diffuse social forces rewarding and penalizing the AI. Another type of incentive method would involve creating a setup wherein the AI can be rewarded and penalized by the project that creates it, and thereby incentivized to act in the interests of the principal. To achieve this, the AI would be placed in a surveillance context that allows its behavior to be monitored and evaluated, either manually or by some automated process. The AI would know that a positive evaluation would bring about some outcome it desires and that a negative evaluation would fail to do so. Theoretically, the reward could be the fulfillment of some convergent instrumental goal. Without knowing anything specific about the AI's motivation system, however, it might be difficult to calibrate the reward mechanism. For example, we could end up with an AI that would be willing to take extreme risks for the sake of a small chance of eventually obtaining control of a large share of the universe. It could be expensive to offer the AI a higher expected utility as reward for cooperation than the AI could hope to achieve by defecting and trying to escape.[10]

A better alternative might be to combine the incentive method with the use of motivation selection to give the AI a final goal that makes it easier to control. Suppose that an AI were designed to have as its final goal that a particular red button inside a command bunker never be pressed. Since the pressing of the button is disvalued intrinsically and not because of its causal consequences, the button could be completely inert: it could be made of Play-Doh. Furthermore, it is irrelevant whether the AI can ever know whether the button had been pressed. What is essential is that the AI believes that the button will more likely remain unpressed if the AI continuously acts in the principal's interest than if it rebels.

Refinements to this setup are possible. Instead of trying to endow an AI with a final goal that refers to a physical button, one could build an AI that places final value on receiving a stream of "cryptographic reward tokens."[11] These would be sequences of numbers serving as keys to ciphers that would have been generated before the AI was created and that would have been built into its motivation system.[12] These special number sequences would be extremely desirable to the AI, constituting a special kind of reward token that the AI could not attain though wireheading.[13] The keys would be stored in a secure location where they could be quickly destroyed if the AI ever made an attempt to seize them. So long as the AI cooperates, the keys are doled out at a steady rate.

The AI could then be given a bounded utility function, and matters could be arranged so that the AI would get (for instance) 99% of its maximum utility from the first reward token; 99% of its remaining utility potential from the second reward token; and so on. By keeping the "discount rate" high, and

by transmitting reward tokens to the AI at a reasonably high rate, one would dissuade the AI from attempting to carry out long-term plans that would increase the expected number of reward tokens in the distal future at the expense of some risk that its reward stream will be cut short or that the next reward token will be cancelled and destroyed. If the AI always believes that defecting (i.e. doing anything other than its best to look after the principal's interests) would incur at least a 2% risk of forfeiting the next reward token, whereas cooperating would reduce the risk of forfeiting the next reward token to below 1%, then a utility-maximizing AI would always cooperate.

What might go wrong with such an incentive scheme? One possibility is that the AI will not trust the human operator to deliver the promised rewards. The track record of human reliability is something other than a straight line of unerring perfection. The AI would reasonably worry that the operator will change his mind, or raise the performance bar, or fail to recognize that the AI has done its part. The AI would also worry about the operator becoming incapacitated. The combined risk of such failures might exceed the risk of attempting to seize control of the reward mechanism. Even a boxed AI possessing the panoply of superpowers is a strong force. (For an AI that is *not* boxed to begin with, hijacking the human-governed reward mechanism may be like taking candy from a baby.)

Another problem with the incentive scheme is that it presupposes that we can tell whether the outcomes produced by the AI are in our interest. … this presupposition is not innocuous.

A full assessment of the feasibility of incentive methods would also have to take into account a range of other factors, including some esoteric considerations that might conceivably make such methods more viable than a preliminary analysis would suggest. In particular, the AI may face ineliminable indexical uncertainty if it could not be sure that it does not inhabit a computer simulation (as opposed to "basement-level," non-simulated physical reality), and this epistemic predicament may radically influence the AI's deliberations (see Box **23.1**).

Box 23.1 Anthropic Capture

The AI might assign a substantial probability to its simulation hypothesis, the hypothesis that it is living in a computer simulation. Even today, many AIs inhabit simulated worlds – worlds consisting of geometric line drawings, texts, chess games, or simple virtual realities, and in which the laws of physics deviate sharply from the laws of physics that we believe govern the world of our own experience. Richer and more complicated virtual worlds will become feasible with improvements in programming techniques and computing power. A mature superintelligence could create virtual worlds

that appear to its inhabitants much the same as our world appears to us. It might create vast numbers of such worlds, running the same simulation many times or with small variations. The inhabitants would not necessarily be able to tell whether their world is simulated or not; but if they are intelligent enough they could consider the possibility and assign it some probability. In light of the simulation argument (a full discussion of which is beyond the scope of this book) that probability could be substantial.[14]

This predicament especially afflicts relatively early-stage superintelligences, ones that have not yet expanded to take advantage of the cosmic endowment. An early-stage superintelligence, which uses only a small fraction of the resources of a single planet, would be much less expensive to simulate than a mature intergalactic superintelligence. Potential simulators – that is, other more mature civilizations – would be able to run great numbers of simulations of such early-stage AIs even by dedicating a minute fraction of their computational resources to that purpose. If at least some (non-trivial fraction) of these mature superintelligent civilizations choose to use this ability, early-stage AIs should assign a substantial probability to being in a simulation.

How an AI would be affected by the simulation hypothesis depends on its values.[15] Consider first an AI that has a "resource-insatiable" final goal, such as the goal of maximizing a utility function that is linear in the number of paperclips caused by the AI to exist – twice as many paperclips, twice as good. Such an AI might care less about the simulation hypothesis, on grounds that its ability to influence how many paperclips are produced looks much more limited if the AI is in a simulation than if it is in basement-level physical reality. The AI might accordingly behave much *as if* it were certain that the simulation hypothesis is false, even if it assigned a fairly high probability to the simulation hypothesis being true.[16] The AI would reason that it does not much matter how it acts if the simulation hypothesis is true, whereas its actions make a great difference if the hypothesis is false; so it would select actions mainly by reference to how good they are conditional on the simulation hypothesis being false.

By contrast, consider an AI that has a more modest final goal, one that could be satisfied with a small amount of resources, such as the goal of receiving some pre-produced cryptographic reward tokens, or the goal of causing the existence of forty-five virtual paperclips. Such an AI should not discount those possible worlds in which it inhabits a simulation. A substantial portion of the AI's total expected utility might derive from those possible worlds. The decision-making of an AI with goals that are easily resource-satiable may therefore – if it assigns a

high probability to the simulation hypothesis – be dominated by considerations about which actions would produce the best result if its perceived world is a simulation. Such an AI (even if it is, in fact, not in a simulation) might therefore be heavily influenced by its beliefs about which behaviors would be rewarded in a simulation.

In particular, if an AI with resource-satiable final goals believes that in most simulated worlds that match its observations it will be rewarded if it cooperates (but not if it attempts to escape its box or contravene the interests of its creator) then it may choose to cooperate. We could therefore find that even an AI with a decisive strategic advantage, one that could in fact realize its final goals to a greater extent by taking over the world than by refraining from doing so, would nevertheless balk at doing so.

> Thus Conscience does make Cowards of us all,
> And thus the Native hue of Resolution
> Is sicklied o'er, with the pale cast of Thought,
> And enterprises of great pith and moment,
> With this regard their Currents turn away,
> And lose the name of Action.
>
> (Shakespeare, *Hamlet*, Act III. Sc. i)

A mere line in the sand, backed by the clout of a nonexistent simulator, could prove a stronger restraint than a two-foot-thick solid steel door.[17]

Stunting

Another possible capability control method is to limit the system's intellectual faculties or its access to information. This might be done by running the AI on hardware that is slow or short on memory. In the case of a boxed system, information inflow could also be restricted.

Stunting an AI in these ways would limit its usefulness. The method thus faces a dilemma: too little stunting, and the AI might have the wit to figure out some way to make itself more intelligent (and thence to world domination); too much, and the AI is just another piece of dumb software. A radically stunted AI is certainly safe but does not solve the problem of how to achieve a controlled detonation: an intelligence explosion would remain possible and would simply be triggered by some other system instead, perhaps at a slightly later date.

One might think it would be safe to build a superintelligence provided it is only given data about some narrow domain of facts. For example, one might build an AT that lacks sensors and that has preloaded into its memory only facts about petroleum engineering or peptide chemistry. But if the AI is

superintelligent – if it is has a superhuman level of *general* intelligence – such data deprivation does not guarantee safety.

There are several reasons for this. First, the notion of information being "about" a certain topic is generally problematic. Any piece of information can in principle be relevant to any topic whatsoever, depending on the background information of a reasoner.[18] Furthermore, a given data set contains information not only about the domain from which the data was collected but also about various circumstantial facts. A shrewd mind looking over a knowledge base that is nominally about peptide chemistry might infer things about a wide range of topics. The fact that certain information is included and other information is not could tell an AI something about the state of human science, the methods and instruments available to study peptides, the fabrication technologies used to make these instruments, and the nature of the brains and societies that conceived the studies and the instruments. It might be that a *superintelligence* could correctly surmise a great deal from what seem, to dull-witted human minds, meager scraps of evidence. Even without any designated knowledge base at all, a sufficiently superior mind might be able to learn much by simply introspecting on the workings of its own psyche – the design choices reflected in its source code, the physical characteristics of its circuitry.[19] Perhaps a super-intelligence could even deduce much about the likely properties of the world *a priori* (combining logical inference with a probability prior biased toward simpler worlds, and a few elementary facts implied by the superintelligence's existence as a reasoning system). It might imagine the consequences of different possible laws of physics: what kind of planets would form, what kind of intelligent life would evolve, what kind of societies would develop, what kind of methods to solve the control problem would be attempted, how those methods could be defeated.[20]

For these reasons it would be imprudent to rely on information deprivation as the main check on a superintelligence's power. Some forms of stunting however, including limiting hardware capacity and preventing the development of specific cognitive faculties, might find use as auxiliary control methods in the developmental phase. They could be used, for instance, to regulate the rate of capacity gain in an intermediary AI and to allow tests to be run at different levels of capability short of full-throttled superintelligence.

Tripwires

A tripwire is a mechanism that performs diagnostic tests on the system (possibly without its knowledge) and effects a shutdown if it detects signs of dangerous activity. Programmers can then scrutinize recordings of the suspicious activity at their leisure, and make changes to the AI's code base or environment before restarting it. If they find that the problem is sufficiently deep-rooted, they could decide to abort the entire project.

Tripwires differ from incentive methods in that they do not rely on the system being aware of the consequences of engaging in forbidden activities. Tripwires are more closely related to stunting methods. Like stunting, tripwires could be used as a temporary safeguard, providing a degree of protection during the development phase. In principle, tripwires can also be used during the operational phase, particularly for a boxed system. However, the ability of tripwires to constrain a full-fledged superintelligence must remain very much in doubt, since it would be hard for us to assure ourselves that such an agent could not find ways to subvert any tripwire devised by the human intellect.

As an adjunct safety measure in the development phase, though, tripwires are important. It would be good practice for the AI's designers to codify their expectations about how the system is supposed to function, and then to circumscribe the AI with tripwires that would be activated if any of these expectations is violated. To make the tests more stringent, "honeypots" could be strategically placed to create temptations for a malfunctioning AI to commit some easily observable violation. For instance, if an AI has been designed in such a way that it is supposed not to want to access the internet, a fake Ethernet port could be installed (leading to an automatic shutdown switch) just to see if they AI tries to use it. (Some examples of tripwires are given in Table 23.1.)

It should be emphasized that the value of a tripwire depends not only on the mechanism itself but also – critically – on how a project reacts when a tripwire is triggered. If the programmers or project managers, impatient to make progress, simply switch the system back on again – or if they do so after making some token modification to prevent the tripwire being triggered on the next run – then no safety has been gained even if the tripwire itself works exactly as intended.

Motivation Selection Methods

Motivation selection methods seek to prevent undesirable outcomes by shaping what the superintelligence wants to do. By engineering the agent's motivation system and its final goals, these methods would produce a superintelligence that would not *want* to exploit a decisive strategic advantage in a harmful way. Since a superintelligent agent is skilled at achieving its ends, if it prefers not to cause harm (in some appropriate sense of "harm") then it would tend not to cause harm (in that sense of "harm").

Motivation selection can involve explicitly formulating a goal or set of rules to be followed (*direct specification*) or setting up the system so that it can discover an appropriate set of values for itself by reference to some implicitly or indirectly formulated criterion (*indirect normativity*). One option in motivation selection is to try to build the system so that it would

Table 23.1 Different kinds of tripwires

Type of monitoring	*Description and examples*
Behavior	Detectors could be placed around a boxed AI to detect attempts to breach the containment. For example, detectors could intercept attempts at radio communication or at accessing internal computational resources intended to be off limits. An "Ethernet port of Eden" could be installed: an apparent connection to the internet that leads to a shutdown switch.
Ability	Automated capability testing could be performed at frequent intervals to determine the AI's skill in various domains. If either the rate of improvement is unexpectedly high or the AI attains a level of competence that brings it close to the potential danger zone, the system could be slowed down or paused, in order to allow the programmers to monitor further developments more closely.
Content	The AI's internal processes could be monitored for suspicious activity. For example, any change to the AI's representation of its final values might trigger an automatic shutdown and review. The AI's resource use and algorithmic methods may also be also be monitored for unexpected patterns. More ambitiously, the AI's beliefs, plans, subgoals, and justifications may be continuously scrutinized, to ensure they conform to the programmers' expectations. Such monitoring of the AI's intentional states would enable a particularly powerful form of content monitoring: the scanning of the AI's cognitive processes for a *conception of deception* – that vulnerable moment when an unfriendly AI first forms the intention to conceal its true intentions.[a] Content monitoring that requires that the AI's intentional states be transparent to the programmers or to an automatic monitoring mechanism may not be feasible for all kinds of AI architectures. (Some neural networks, for instance, are opaque, as they represent information holistically and in ways that do not necessarily match up with human concepts.) This may be a reason to avoid using such architectures.

[a] The moment after the conception of deception, the AI might contrive to erase the trace of its mutinous thought. It is therefore important that this tripwire operate continuously. It would also be good practice to use a "flight recorder" that stores a complete trace of all the AI's activity (including exact timing of keyboard input from the programmers), so that its trajectory can be retraced or analyzed following an automatic shutdown. The information could be stored on a write-once-read-many medium.

have modest, non-ambitious goals (*domesticity*). An alternative to creating a motivation system from scratch is to select an agent that already has an acceptable motivation system and then augment that agent's cognitive powers to make it superintelligent, while ensuring that the motivation system does not get corrupted in the process (*augmentation*). Let us look at these in turn.

Direct Specification

Direct specification is the most straightforward approach to the control problem. The approach comes in two versions, rule-based and consequentialist, and involves trying to explicitly define a set of rules or values that will cause even a free-roaming superintelligent AI to act safely and beneficially. Direct specification, however, faces what may be insuperable obstacles, deriving from both the difficulties in determining which rules or values we would wish the AI to be guided by and the difficulties in expressing those rules or values in computer-readable code.

The traditional illustration of the direct rule-based approach is the "three laws of robotics" concept, formulated by science fiction author Isaac Asimov in a short story published in 1942.[21] The three laws were: (1) A robot may not injure a human being or, through inaction, allow a human being to come to harm; (2) A robot must obey any orders given to it by human beings, except where such orders would conflict with the First Law; (3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Law. Embarrassingly for our species, Asimov's laws remained state-of-the-art for over half a century: this despite obvious problems with the approach, some of which are explored in Asimov's own writings (Asimov probably having formulated the laws in the first place precisely so that they would fail in interesting ways, providing fertile plot complications for his stories).[22]

Bertrand Russell, who spent many years working on the foundations of mathematics, once remarked that "everything is vague to a degree you do not realize till you have tried to make it precise."[23] Russell's dictum applies in spades to the direct specification approach. Consider, for example, how one might explicate Asimov's first law. Does it mean that the robot should minimize the probability of any human being coming to harm? In that case the other laws become otiose since it is always possible for the AI to take some action that would have at least some microscopic effect on the probability of a human being coming to harm. How is the robot to balance a large risk of a few humans coming to harm versus a small risk of many humans being harmed? How do we define "harm" anyway? How should the harm of physical pain be weighed against the harm of architectural ugliness or social injustice? Is a sadist harmed if he is prevented from

tormenting his victim? How do we define "human being"? Why is no consideration given to other morally considerable beings, such as sentient nonhuman animals and digital minds? The more one ponders, the more the questions proliferate.

Perhaps the closest existing analog to a rule set that could govern the actions of a superintelligence operating in the world at large is a legal system. But legal systems have developed through a long process of trial and error, and they regulate relatively slowly changing human societies. Laws can be revised when necessary. Most importantly, legal systems are administered by judges and juries who generally apply a measure of common sense and human decency to ignore logically possible legal interpretations that are sufficiently obviously unwanted and unintended by the lawgivers. It is probably humanly impossible to explicitly formulate a highly complex set of detailed rules, have them apply across a highly diverse set of circumstances, and get it right on the first implementation.[24]

Problems for the direct consequentialist approach are similar to those for the direct rule-based approach. This is true even if the AI is intended to serve some apparently simple purpose such as implementing a version of classical utilitarianism. For instance, the goal "Maximize the expectation of the balance of pleasure over pain in the world" may appear simple. Yet expressing it in computer code would involve, among other things, specifying how to recognize pleasure and pain. Doing this reliably might require solving an array of persistent problems in the philosophy of mind – even just to obtain a correct account expressed in a natural language, an account which would then, somehow, have to be translated into a programming language.

A small error in either the philosophical account or its translation into code could have catastrophic consequences. Consider an AI that has hedonism as its final goal, and which would therefore like to tile the universe with "hedonium" (matter organized in a configuration that is optimal for the generation of pleasurable experience). To this end, the AI might produce computronium (matter organized in a configuration that is optimal for computation) and use it to implement digital minds in states of euphoria. In order to maximize efficiency, the AI omits from the implementation any mental faculties that are not essential for the experience of pleasure, and exploits any computational shortcuts that according to its definition of pleasure do not vitiate the generation of pleasure. For instance, the AI might confine its simulation to reward circuitry, eliding faculties such as memory, sensory perception, executive function, and language; it might simulate minds at a relatively coarse-grained level of functionality, omitting lower-level neuronal processes; it might replace commonly repeated computations with calls to a lookup table; or it might put in place some arrangement whereby multiple minds would share most parts of their underlying computational machinery (their "supervenience bases" in

philosophical parlance). Such tricks could greatly increase the quantity of pleasure producible with a given amount of resources. It is unclear how desirable this would be. Furthermore, if the AI's criterion for determining whether a physical process generates pleasure is wrong, then the AI's optimizations might throw the baby out with the bathwater: discarding something which is inessential according to the AI's criterion yet essential according to the criteria implicit in our human values. The universe then gets filled not with exultingly heaving hedonium but with computational processes that are unconscious and completely worthless – the equivalent of a smiley-face sticker xeroxed trillions upon trillions of times and plastered across the galaxies.

Domesticity

One special type of final goal which might be more amenable to direct specification than the examples given above is the goal of self-limitation. While it seems extremely difficult to specify how one would want a superintelligence to behave in the world *in general* – since this would require us to account for all the trade-offs in all the situations that could arise – it might be feasible to specify how a super-intelligence should behave in one particular situation. We could therefore seek to motivate the system to confine itself to acting on a small scale, within a narrow context, and through a limited set of action modes. We will refer to this approach of giving the AI final goals aimed at limiting the scope of its ambitions and activities as "domesticity."

For example, one could try to design an AI such that it would function as a question-answering device (an "oracle" …). Simply giving the AI the final goal of producing maximally accurate answers to any question posed to it would be unsafe. … (Reflect …. that this goal would incentivize the AI to take actions to ensure that it is asked easy questions.) To achieve domesticity, one might try to define a final goal that would somehow overcome these difficulties: perhaps a goal that combined the desiderata of answering questions correctly and minimizing the AI's impact on the world except whatever impact results as an incidental consequence of giving accurate and non-manipulative answers to the questions it is asked.[25]

The direct specification of such a domesticity goal is more likely to be feasible than the direct specification of either a more ambitious goal or a complete rule set for operating in an open-ended range of situations. Significant challenges nonetheless remain. Care would have to be taken, for instance, in the definition of what it would be for the AI to "minimize its impact on the world" to ensure that the measure of the AI's impact coincides with our own standards for what counts as a large or a small impact. A bad measure would lead to bad trade-offs. There are also other kinds of risk associated with building an oracle, which we will discuss later.

There is a natural fit between the domesticity approach and physical containment. One would try to "box" an AI such that the system is *unable* to escape while simultaneously trying to shape the AI's motivation system such that it would be *unwilling* to escape even if it found a way to do so. Other things equal, the existence of multiple independent safety mechanisms should shorten the odds of success.[26]

Indirect Normativity

If direct specification seems hopeless, we might instead try indirect normativity. The basic idea is that rather than specifying a concrete normative standard directly, we specify a process for deriving a standard. We then build the system so that it is motivated to carry out this process and to adopt whatever standard the process arrives at.[27] For example, the process could be to carry out an investigation into the empirical question of what some suitably idealized version of us would prefer the AI to do. The final goal given to the AI in this example could be something along the lines of "achieve that which we would have wished the AI to achieve if we had thought about the matter long and hard."

… Indirect normativity is a very important approach to motivation selection. Its promise lies in the fact that it could let us offload to the superintelligence much of the difficult cognitive work required to carry out a direct specification of an appropriate final goal.

Augmentation

The last motivation selection method on our list is augmentation. Here the idea is that rather than attempting to design a motivation system *de novo,* we start with a system that already has an acceptable motivation system, and enhance its cognitive faculties to make it superintelligent. If all goes well, this would give us a superintelligence with an acceptable motivation system.

This approach, obviously, is unavailing in the case of a newly created seed AI. But augmentation is a potential motivation selection method for other paths to superintelligence, including brain emulation, biological enhancement, brain–computer interfaces, and networks and organizations, where there is a possibility of building out the system from a normative nucleus (regular human beings) that already contains a representation of human value.

The attractiveness of augmentation may increase in proportion to our despair at the other approaches to the control problem. Creating a motivation system for a seed AI that remains reliably safe and beneficial under recursive self-improvement even as the system grows into a mature superintelligence

is a tall order, especially if we must get the solution right on the first attempt. With augmentation, we would at least start with a system that has familiar and human-like motivations.

On the downside, it might be hard to ensure that a complex, evolved, kludgy, and poorly understood motivation system, like that of a human being, will not get corrupted when its cognitive engine blasts into the stratosphere. As discussed earlier, an imperfect brain emulation procedure that preserves intellectual functioning may not preserve all facets of personality. The same is true (though perhaps to a lesser degree) for biological enhancements of cognition, which might subtly affect motivation, and for collective intelligence enhancements of organizations and networks, which might adversely change social dynamics (e.g. in ways that debase the collective's attitude toward outsiders or toward its own constituents). If superintelligence is achieved via any of these paths, a project sponsor would find guarantees about the ultimate motivations of the mature system hard to come by. A mathematically well-specified and foundationally elegant AI architecture might – for all its non-anthropomorphic otherness – offer greater transparency, perhaps even the prospect that important aspects of its functionality could be formally verified.

In the end, however one tallies up the advantages and disadvantages of augmentation, the choice as to whether to rely on it might be forced. If superintelligence is first achieved along the artificial intelligence path, augmentation is not applicable. Conversely, if superintelligence is first achieved along some non-AI path, then many of the other motivation selection methods are inapplicable. Even so, views on how likely augmentation would be to succeed do have strategic relevance insofar as we have opportunities to influence which technology will first produce superintelligence.

Synopsis

A quick synopsis might be called for before we close this chapter. We distinguished two broad classes of methods for dealing with the agency problem at the heart of AI safety: capability control and motivation selection. Table 23.2 gives a summary.

Each control method comes with potential vulnerabilities and presents different degrees of difficulty in its implementation. It might perhaps be thought that we should rank them from better to worse, and then opt for the best method. But that would be simplistic. Some methods can be used in combination whereas others are exclusive. Even a comparatively insecure method may be advisable if it can easily be used as an adjunct, whereas a strong method might be unattractive if it would preclude the use of other desirable safeguards.

Table 23.2 Control methods

Capability control	
Boxing methods	The system is confined in such a way that it can affect the external world only through some restricted, pre-approved channel. Encompasses physical and informational containment methods.
Incentive methods	The system is placed within an environment that provides appropriate incentives. This could involve social integration into a world of similarly powerful entities. Another variation is the use of (cryptographic) reward tokens. 'Anthropic capture" is also a very important possibility but one that involves esoteric considerations.
Stunting	Constraints are imposed on the cognitive capabilities of the system or its ability to affect key internal processes.
Tripwires	Diagnostic tests are performed on the system (possibly without its knowledge) and a mechanism shuts down the system if dangerous activity is detected.
Motivation selection	
Direct specification	The system is endowed with some directly specified motivation system, which might be consequentialist or involve following a set of rules.
Domesticity	A motivation system is designed to severely limit the scope of the agent's ambitions and activities.
Indirect normativity	Indirect normativity could involve rule-based or consequentialist principles, but is distinguished by its reliance on an indirect approach to specifying the rules that are to be followed or the values that are to be pursued.
Augmentation	One starts with a system that already has substantially human or benevolent motivations, and enhances its cognitive capacities to make it superintelligent.

It is therefore necessary to consider what package deals are available. We need to consider what type of system we might try to build, and which control methods would be applicable to each type. [...]

Notes

1. E.g., Laffont and Martimort (2002).
2. Suppose a majority of voters want their country to build some particular kind of superintelligence. They elect a candidate who promises to do their bidding, but

they might find it difficult to ensure that the candidate, once in power, will follow through on her campaign promise and pursue the project in the way that the voters intended. Supposing she is true to her word, she instructs her government to contract with an academic or industry consortium to carry out the work; but again there are agency problems, the bureaucrats in the government department might have their own views about what should be done and may implement the project in a way that respects the letter but not the spirit of the leader's instructions. Even if the government department does its job faithfully, the contracted scientific partners might have their own separate agendas. The problem recurs on many levels. The director of one of the participating laboratories might lie awake worrying about a technician introducing an unsanctioned element into the design – imagining Dr. T. R. Eason sneaking into his office late one night, logging into the project code base, rewriting a part of the seed AI's goal system. Where it was supposed to say "serve humanity," it now says "serve Dr. T. R. Eason."

3. Even for superintelligence development, though, there could be a role for behavioral testing – as one auxiliary element within a wider battery of safety measures. Should an AI misbehave in its developmental phase, something is clearly awry – though, importantly, the converse does not hold.
4. In a classic exploit from 1975, Steven Dompier wrote a program for the Altair 8800 that took advantage of this effect (and the absence of shielding around the microcomputer's case). Running the program caused the emission of electromagnetic waves that would produce music when one held a transistor radio close to the computer (Driscoll 2012). The young Bill Gates, who attended a demo, reported that he was impressed and mystified by the hack (Gates 1975). There are in any case plans to design future chips with built-in Wi-Fi capabilities (Greene 2012).
5. It is no light matter to have held a conviction, which, had we had an opportunity to act upon it, could have resulted in the ruination of all our cosmic endowment. Perhaps one could argue for the following principle: if somebody has in the past been certain on N occasions that a system has been improved sufficiently to make it safe, and each time it was revealed that they were wrong, then on the next occasion they are not entitled to assign a credence greater than $1/(N + 1)$ to the system being safe.
6. In one informal experiment, the role of the AI was played by an intelligent human. Another individual played the role of gatekeeper and was tasked with not letting the AI out of the box. The AI could communicate with the gatekeeper only by text and was given two hours to persuade the gatekeeper to let it out. In three cases out of five, with different individuals playing the gatekeeper, the AI escaped (Yudkowsky 2002). What a human can do, a superintelligence can do too. (The reverse, of course, does not hold. Even if the task for a real superintelligence were harder – maybe the gatekeepers would be more strongly motivated to refrain from releasing the AI than the individuals playing gatekeeper in the experiment – the superintelligence might still succeed where a human would fail.)
7. One should not overstate the marginal amount of safety that could be gained in this way. Mental imagery can substitute for graphical display. Consider the impact books can have on people – and books are not even interactive.

8. See also Chalmers (2010). It would be a mistake to infer from this that there is *no* possible use in building a system that will never be observed by any outside entity. One might place a final value on what goes on inside such a system. Also, other people might have preferences about what goes on inside such a system, and might therefore be influenced by its creation or the promise of its creation. Knowledge of the existence of certain kinds of isolated systems (ones containing observers) can also induce anthropic uncertainty in outside observers, which may influence their behavior.
9. One might wonder why social integration is considered a form of capability control. Should it not instead be classified as a motivation selection method on the ground that it involves seeking to influence a system's behavior by means of incentives? We will look closely at motivation selection presently; but, in answer to this question, we are construing motivation selection as a cluster of control methods that work by selecting or shaping a system's final goals – goals sought for their own sakes rather than for instrumental reasons. Social integration does not target a system's final goals, so it is not motivation selection. Rather, social integration aims to limit the system's effective capabilities: it seeks to render the system incapable of achieving a certain set of outcomes – outcomes in which the system attains the benefits of defection without suffering the associated penalties (retribution, and loss of the gains from collaboration). The hope is that by limiting which outcomes the system is able to attain, the system will find that the most effective remaining means of realizing its final goals is to behave cooperatively.
10. This approach may be somewhat more promising in the case of an emulation believed to have anthropomorphic motivations.
11. I owe this idea to Carl Shulman.
12. Creating a cipher certain to withstand a superintelligent code-breaker is a nontrivial challenge. For example, traces of random numbers might be left in some observer's brain or in the microstructure of the random generator, from whence the superintelligence can retrieve them; or, if pseudorandom numbers are used, the superintelligence might guess or discover the seed from which they were generated. Further, the superintelligence could build large quantum computers, or even discover unknown physical phenomena that could be used to construct new kinds of computers.
13. The AI could wire itself to *believe* that it had received a reward tokens, but this should not make it wirehead if it is designed to want the reward tokens (as opposed to wanting to be in a state in which it has certain beliefs about the reward tokens).
14. For the original article, see Bostrom (2003). See also Elga (2004).
15. Shulman (2010).
16. Basement-level reality presumably contains more computational resources than simulated reality, since any computational processes occurring in a simulation are also occurring on the computer running the simulation. Basement-level reality might also contain a wealth of other physical resources which could be hard for simulated agents to access – agents that exist only at the indulgence of powerful simulators who may have other uses in mind for those resources.

(Of course, the inference here is not strictly deductively valid: in principle, it could be the case that universes in which simulations are run contain so much more resources that simulated civilizations on average have access to more resources than non-simulated civilizations, even though each non-simulated civilization that runs simulations has more resources than all the civilizations it simulates do combined.)

17. There are various further esoteric considerations that might bear on this matter, the implications of which have not yet been fully analyzed. These considerations may ultimately be crucially important in developing an all-things--considered approach to dealing with the prospect of an intelligence explosion. However, it seems unlikely that we will succeed in figuring out the practical import of such esoteric arguments unless we have first made some progress on the more mundane kinds of consideration that are the topic of most of this book.
18. Cf., e.g., Quine and Ullian (1978).
19. Which an AI might investigate by considering the performance characteristics of various basic computational functionalities, such as the size and capacity of various data buses, the time it takes to access different parts of memory, the incidence of random bit flips, and so forth.
20. Perhaps the prior could be (a computable approximation of) the Solomonoff prior, which assigns probability to possible worlds on the basis of their algorithmic complexity. See Li and Vitányi (2008).
21. Asimov (1942). To the three laws were later added a "Zeroth Law": "(0) A robot may not harm humanity, or, by inaction, allow humanity to come to harm" (Asimov 1985).
22. Cf. Gunn (1982).
23. Russell (1986, 161f).
24. Similarly, although some philosophers have spent entire careers trying to carefully formulate deontological systems, new cases and consequences occasionally come to light that necessitate revisions. For example, deontological moral philosophy has in recent years been reinvigorated through the discovery of a fertile new class of philosophical thought experiments, "trolley problems," which reveal many subtle interactions among our intuitions about the moral significance of the acts/omissions distinction, the distinction between intended and unintended consequences, and other such matters; see, e.g., Kamm (2007).
25. Armstrong (2010).
26. As a rule of thumb, if one plans to use multiple safety mechanisms to contain an AI, it may be wise to work on each one *as if* it were intended to be the sole safety mechanism and *as if* it were therefore required to be individually sufficient. It one puts a leaky bucket inside another leaky bucket, the water still comes out.
27. A variation of the same idea is to build the AI so that it is continuously motivated to act on its best guesses about what the implicitly defined standard is. In this setup, the AI's final goal is always to act on the implicitly defined standard, and it pursues an investigation into what this standard is only for instrumental reasons.

References

Armstrong, Stuart. 2010. *Utility Indifference*. Technical Report 2010-1. Oxford: Future of Humanity Institute, University of Oxford.

Asimov, Isaac. 1942. "Runaround." *Astounding Science-Fiction*, March, 94–103.

Asimov, Isaac. 1985. *Robots and Empire*. New York: Doubleday.

Bostrom, Nick. 2003. "Are We Living in a Computer Simulation?" *Philosophical Quarterly* **53** (211): 243–255.

Chalmers, David John. 2010. "The Singularity: A Philosophical Analysis." *Journal of Consciousness Studies* **17** (9–10): 7–65.

Driscoll, Kevin. 2012. "Code Critique:' Altair Music of a Sort.'" Paper presented at Critical Code Studies Working Group Online Conference, 2012, February 6.

Elga, Adam. 2004. "Defeating Dr. Evil with Self-Locating Belief." *Philosophy and Phenomenological Research* **69** (2): 383–396.

Gates, Bill. 1975. "Software Contest Winners Announced." *Computer Notes* **1**(2): 1.

Greene, Kate. 2012. "Intel's Tiny Wi-Fi Chip Could Have a Big Impact." *MIT Technology Review*, September 21.

Gunn, James E. 1982. *Isaac Asimov: The Foundations of Science Fiction*. Science-Fiction Writers. New York: Oxford University Press.

Kamm, Frances M. 2007. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. Oxford Ethics Series. New York: Oxford University Press.

Laffont, Jean-Jacques, and Martimort, David. 2002. *The Theory of Incentives: The Principal- Agent Model*. Princeton, NJ: Princeton University Press.

Li, Ming, and Vitányi, Paul M. B. 2008. *An Introduction to Kolmogorov Complexity and Its Application*. Texts in Computer Science. New York: Springer.

Quine, Willard Van Orman, and Ullian, Joseph Silbert. 1978. *The Web of Belief*, ed. Richard Malin Ohmann, vol. **2**. New York: Random House.

Russell, Bertrand. 1986. "The Philosophy of Logical Atomism." In *The Philosophy of Logical Atomism and Other Essays 1914–1919*, ed. John G. Slater, **8**: 157–244. The Collected Papers of Bertrand Russell. Boston: Allen & Unwin.

Shulman, Carl. 2010.*Omohundro's "Basic AI Drives" and Catastrophic Risks*. San Francisco, CA: Machine Intelligence Research Institute.

Yudkowsky, Eliezer. 2002. "The AI-Box Experiment." Retrieved January 15, 2012. Available at http://yudkowsky.net/singularity/aibox.