Thinking Inside the Box: Controlling and Using an Oracle AI

Stuart Armstrong
Anders Sandberg
Nick Bostrom
[(2012) *Minds and Machines*, forthcoming]
www.nickbostrom.com

Abstract

There is no strong reason to believe that human-level intelligence represents an upper limit of the capacity of artificial intelligence, should it be realized. This poses serious safety issues, since a superintelligent system would have great power to direct the future according to its possibly flawed motivation system. Solving this issue in general has proven to be considerably harder than expected. This paper looks at one particular approach, Oracle AI. An Oracle AI is an AI that does not act in the world except by answering questions. Even this narrow approach presents considerable challenges. In this paper, we analyse and critique various methods of controlling the AI. In general an Oracle AI might be safer than unrestricted AI, but still remains potentially dangerous.

Keywords: Artificial Intelligence, Superintelligence, Security, Risks, Motivational control, Capability control

1 Introduction

There are many motivations to pursue the goal of artificial intelligence (AI). While some motivations are non-instrumental, such as scientific and philosophical curiosity about the nature of thinking or a desire for creating non-human beings, a strong set of motivations is the instrumental utility of AI. Such machines would benefit their owners by being able to do tasks that currently require human intelligence, and possibly tasks that are beyond human intelligence. From an economic perspective the possibility of complementing or substituting expensive labour with cheaper software promises very rapid growth rates and high productivity (Hanson 2001) (Kaas, et al. 2010). The introduction of sufficiently advanced AI would have profound effects on most aspects of society, making careful foresight important.

While most considerations about the mechanisation of labour have focused on AI with intelligence up to the human level, there is no strong reason to believe humans represent an upper limit of possible intelligence. The human brain has evolved under various biological constraints (e.g. food availability, birth canal size, trade-offs with other organs, the requirement of using biological materials) that do not exist for an artificial system. Besides having different hardware, an AI might employ more effective algorithms that cannot be implemented well in the human cognitive architecture (e.g. making use of very large and exact working memory, stacks, mathematical modules or numerical simulation) or employ tricks that are not feasible for humans, such as running multiple instances whose memories and conclusions are eventually merged. In addition, if an AI system possesses sufficient abilities, it would be able to assist in developing better AI. Since AI development is an expression of human intelligence, at least some AI might achieve this form of intelligence, and beyond a certain point would accelerate the development far beyond the current rate (Chalmers 2010) (Kurzweil 2005) (Bostrom 2004).

While the likelihood of superintelligent AI is hotly debated, the mere possibility raises worrisome policy questions. In this paper, we are taking intelligence to mean mainly the cognitive ability to achieve goals. Hence we should expect superintelligent systems to be significantly better at achieving their goals than humans (Yudkowsky 2008). This produces a risky power differential. The appearance of superintelligence appears to pose an existential risk: a possibility that humanity is annihilated or has its potential drastically curtailed indefinitely (Bostrom 2001). This could come about through a number of ways: by enabling self-reinforcing systems that limit human potential (e.g. a global police state (Caplan 2008)), by out-competing humans or human values (see the 'mindless outsourcers' in (Bostrom 2004)), or by acting with great power in such a malevolent or indifferent manner that humanity goes extinct. The last possibility could occur due to badly formulated initial motivations or goals, or gradual evolution towards human-unfriendly behaviours (Omohundro 2008).

In the space of possible motivations[1], likely a very small fraction is compatible with coexistence with humans[2]. A randomly selected motivation can hence be expected to be dangerous. A simple example is that of a paperclip maximiser: an AI with a utility function that seeks to maximize the number of paperclips. This goal is a not too implausible test command for a new system, yet it would result in an AI willing to sacrifice the world and everyone in it, to make more paperclips. If superintelligent, the AI would be very good at converting the world into paperclips, even if it realizes its creators actually didn't want that many paperclips – but stopping would lead to fewer paperclips, and that's what its goal is (Yudkowsky 2009) (Bostrom 2003b). This is a common problem in computer programming, with the program going beyond the implicit bounds of what was expected of it (or users giving, in retrospect, mistaken commands). So unless valuing humans is an integral part of the superintelligence's setup, we can expect that it will see us as mere tools or obstacles for its own goals.

There are several approaches to AI risk. The most common at present is to hope that it is no problem: either sufficiently advanced intelligences will converge towards human-compatible behaviour, a solution will be found closer to the time when they are actually built, or they cannot be built in the first place. While such considerations might turn out to be true, the arguments for them appear relatively uncertain, making it problematic to gamble existential risk only on them.

---

[1] It is generally argued that intelligence and motivation are orthogonal, that a high intelligence is no guarantor of safe motives (Bostrom 2012).

[2] Humans have many preferences – survival, autonomy, hedonistic pleasure, overcoming challenges, satisfactory interactions, and countless others – and we want them all satisfied, to some extent. But a randomly chosen motivation would completely disregard half of these preferences (actually it would disregard much more, as these preferences are highly complex to define – we wouldn't want any of the possible 'partial survival' motivations, for instance).

Another approach is to assume that the behaviour of superintelligent agents will be constrained by other agents on par with them in power, similar to how current humans and organisations are constrained by each other and higher-level institutions (Sandberg 2001). However, this presupposes that the rate of AI development is slow enough that there is time to formulate a cooperative framework and that there will be more than one superintelligent agent. And there is no reason to suppose that a world with several AIs is more amicable to humans than a world with a single AI. Indeed, humans may end up a casualty of the competition, even if none of the AIs would individually want that outcome, because none of them could afford to take steps to protect humans without losing out.

A proactive approach is to attempt to design a 'friendly AI', which is designed to be of low risk[3] (Yudkowsky 2001a) (Yudkowsky 2001b). This might involve safeguards against developing into dangerous directions and top-level goals that include some form of human wellbeing. This approach requires both the conceptualization of sufficient safeguards and the proper implementation of those safeguards in the first AI that achieves superintelligence. It hence depends on developing a workable "friendliness theory" before full superintelligence is achieved, cooperation with the AI developers, and correct implementation. The first requirement is essentially the "inverse morality problem": to construct goals, values, or motivational structures that produce the right class of actions in an agent that is, by assumption, far more intelligent than the person implementing the construction.

When discussing "friendliness", one common suggestion is the Oracle AI (OAI)[4]. The idea is to construct an AI that does not act, but only answers questions. While superintelligent "genies" that try to achieve the wishes of their owners and sovereign AIs that act according to their own goals are obviously dangerous, oracles appear more benign. While owners could potentially use them in selfish or destructive ways – and their answers might in themselves be dangerous (Bostrom 2011) – they do not themselves pose a risk. Or do they? In stories, bottles are rarely the solution to genies.

This paper attempts to analyse the problem of 'boxing' a potentially unfriendly superintelligence. The key question is: are there strategies that reduce the potential existential risk from a superintelligent AI so much that, while implementing it as a free AI would be unsafe, an oracle implementation would be safe?

The paper will start by laying out the general design assumptions for the OAI. Then it will touch upon some of the risks and dangers deriving from the humans running and interacting with the OAI. The central section is section 4, "methods of control", which analyses the various methods for maintaining control over the OAI, and discusses their usefulness and their weaknesses. The final section looks at some of the other problematic issues concerning the OAI (such as its ability to simulate human beings within it and its status as a moral agent itself).

[3] Friendliness should not be interpreted here as social or emotional friendliness, but simply a shorthand for whatever behavioural or motivational constraints that keeps a superintelligent system from deliberately or accidentally harming humans.

[4] Another common term is "AI-in-a-box".

2 Conceptual architecture of the AI

The possible designs of the OAI are innumerable and it is impossible to predict in detail how they will be implemented. For the purposes of this paper, however, we will assume that the OAI's architecture follows this general format:

1) The OAI is implemented in a spatially limited physical substrate, such as a computer.
2) The OAI may be shut off or reset without destroying its physical substrate, and restarted easily.
3) The OAI's background information comes in the form of a separate "read-only" module that may be connected and disconnect as needed.

Most of the present paper still applies to OAIs that do not follow one or more of these restrictions.

The OAI is assumed (when otherwise not noted) to be of human-equivalent intelligence or beyond; less capable systems are unlikely to pose much of a threat.

2.1 Direct programming, self-improvement and evolution

It is not the purpose of this paper to speculate how the OAI will be programmed, but it is important to note the three main approaches under consideration (as well as mixtures between them). The first, seemingly the hardest, is to directly code the entire OAI, just as if it were a traditional program. Another avenue is to start with a 'seed' AI, of limited intelligence but with the ability to self-improve, in the hope that it will transform itself into a much more intelligent entity, and continue doing so again and again using its growing intelligence (Good 1965) (Yudkowsky 2001c). Finally, it might be possible to use directed evolution as a way of constructing an intelligent entity, making different putative AIs compete according to specific criteria, varying them, and choosing the most successful at each generation. This last approach has the advantage that it has worked in the past: we ourselves have evolved to intelligence, so given enough time and resources, so should an AI be evolvable to at least our level.

The different approaches pose different challenges for controlling the resulting OAI. In the first case, the code is clear to us, because we wrote it; the uncertainties are only about what it will result when run in the real world. Even simple programs will often behave in unexpected ways, and it is only subsequently, after carefully parsing the code and its execution, that the programmer determines how the behaviour was written into it from the beginning. An OAI is such an advanced program compared to any today that we wouldn't be able to predict its behaviour simply from reading or even writing its code.

Conversely, however, if the code is obscure to us, this adds an extra layer of uncertainty and complexity. Even if we have access to the physical code, our ability to interpret its meaning or predict its behaviour is even harder when we did not design it ourselves; this challenge is heightened in the self-improving case and extreme in the evolved case, where the code is likely to be incomprehensible to us at every level.

Finally, if the OAI is changing, then the methods of control discussed in section 4 are necessary to ensure not only that the current OAI is safe and accurate, but that the next one will be as well, thus ensuring a continuity of precautions during a controlled intelligence ascent. This is easiest, but also most vital, in the self-improvement case, where the current OAI gets to directly determine the mind of the next OAI.

2.2 Utility functions

One way to conceive an AI's algorithm is by splitting it into two basic components: the intelligence module, capable of making smart decisions but with no intrinsic purpose or direction, and a utility function (von Neumann and Morgenstern 1944), representing the motivational structure of the AI. The AI will then be devoted to maximising its expected utility. The utility function assigns a single number to every possible world and it deals explicitly with probability and uncertainty, making it an ideal format for the AI to work with.

It has been argued that any self-improving AI that can change its motivational structure is likely to move it in the direction of a utility function (Omohundro 2008). If this is the case, it is quite likely that the OAI we are to deal with will have its motivational structure implemented in this form. This is good news, as this division of labour allows us to focus on making the utility function safe, while ignoring the rest of the OAI. Most solutions presented here will not presuppose that the AI's motivational structure is a utility function, but nearly all are improved if it is.

2.3 Accuracy metric: an AI-complete problem

Apart from safety, the most important requirement for an OAI is that it be accurate, to the best of its ability. The whole point of having an 'Oracle' AI is to get useful answers to our questions. "To the best of its ability" means to the best of its current ability; we wouldn't want, for instance, the OAI to attempt to gain control of real-world resources to build a better version of itself that will then answer the question better. Preventing this behaviour is mainly up to our methods of control (see section 4), but the risk is worth bearing in mind when designing the accuracy metric.

Informative accuracy, not the strict truth, is the requirement. Responding to a question on the likely winner of the next election with a detailed list of which atoms will be in which position as a result of that election is not answering the question. The OAI must be motivated to provide human-understandable answers. This, however, will require it both to understand and implement human concepts. Hence accuracy, in general, is an *AI-complete* problem[5] (though narrower AIs dealing with narrower problems do not require full AI-completeness to function effectively). Though accuracy is much easier than friendliness (Yudkowsky 2001b), it does require that the OAI be capable of understanding hard human concepts that are defined only inside our brains.

[5] A term coined by Fanya Montalvo (Mallery 1988) by analogy to the mathematical concept of NP-completeness: a problem is AI-complete if an AI capable of solving it would reasonably also be able to solve all major outstanding problems in AI.

For we have yet to quantify the level of distortion and simplification that is permitted in the OAI's answers, nor have we even rigorously defined the terms "distortion" and "simplification" in a way that non-human entities could understand. Even intuitively, we find these terms ambiguous. And even concepts that we do intuitively understand are already hard to formalise – we lack a precise definition of anger, for instance, through we all 'know' to a large extent what anger is.

If the OAI is wired too much towards truthfulness, it will answer "undetermined" to nearly every question. If it is wired too much towards simplification, it will give a "yes" or "no" answer in situations where a more complicated answer is called for. The correct answer to "have you stopped beating your wife?" is not to choose between "yes" and "no", picking the answer that is slightly more accurate than the other. More complicated concepts – such as requiring the OAI to answer "yes", "no", or "the answer will confuse you" depending on how we would react to the answer – depend on the OAI having a good understanding of "the answer will confuse you" or similar concepts, and still most likely remains an AI-complete problem.

In theory, much could be gained from separating the two issues: to have an OAI dedicated to answering our questions truthfully, and an interpreter AI dedicated to translating the answer into a human-comprehensible format. This does not, however, represent a gain for security: the OAI still has tremendous power to influence us, and the interpreter AI must be able to solve the hard human translation problem, so must be a complex AI itself, with the attendant security risks. Since the OAI must be able to understand human terminology to answer most useful questions, it is probably better to leave the translation problem to it.

A more promising alternative is to have the OAI output its internal probability estimate for a binary outcome: for example, giving a 75% chance of "yes" and 25% chance of "no". The scope for ambiguity is reduced here, though not entirely eliminated: its versions of "yes" and "no" have to agree with ours.

But it is not the purpose of this paper to figure out how to code the accuracy metric, nor solve the translation problem. It is enough to realise it is a hard problem, one that will need to be addressed, probably either through advanced direct coding (similar to motivational rule-based methods, see section 4.2.1) or through training the AI by providing feedback on test questions (similar to motivational black-box methods, see section 4.2.2).

3 Human-level considerations

Crude hackers attack the mechanical and computerised elements of a system. Sophisticated hackers attack the weakest point of a system: the human element. And the human component of an OAI project is a point of exceptional vulnerability. Humans are error-prone, power-hungry, and vulnerable to social manipulation. These weaknesses reinforce each other, and competition (between different individuals in the same OAI project, between different OAI projects, between different countries) will exacerbate all of them.

Humans are very error-prone in most domains (Ord, Hillerbrand and Sandberg 2010) (Kahneman, Slovic and Tversky 1982), and the sort of cautious, patient measures needed to deal with an OAI are alien to our nature. A bureaucratic implementation, with everything requiring many specific, precisely defined steps before anything can happen may help to mitigate these problems. Humans, however, are skilled at working around bureaucracies, so it may be necessary to automate most of these steps to remove them from direct human control.

But the greatest potential errors are conceptual, not just mistakes or oversights. If the OAI is created without many levels of precautions, or if these precautions are badly designed, a catastrophe is possible. And there are many human biases – overconfidence, positive outcome bias, status quo bias, etc. – that make the need for these precautions less obvious, and hence less likely to be implemented when the time comes. AI research has been going on for a long time (McCarthy, et al. 1956), was absurdly over-confident in the imminence of AI for an extended period (Simon 1965) (Russell and Norvig 2009), but researchers have paid scant attention to safety until recently (Bostrom 2000). Indeed, science fiction works have given much more, and earlier (Asimov 1942), attention to the issue of safety than serious academic works have. This is the clearest indication that the OAI designers will, if left to their own devices, most likely neglect security. If advanced AI had been possible at any time in the last half-century, it would have had high chances of ending in disaster. And though safety has been more prominent recently, there are strong competitive pressures towards rushing the job.

Domain-limited oracles are less of a risk. If the OAI has great skills in solving the Navier-Stokes equations, but little else, then it poses correspondingly fewer risks. OAIs that are restricted to narrow scientific or technical questions, if such a restriction could be verified, are relatively safe. However, the task of advancing any field involving human-centred issues – such as economics, marketing, politics, language understanding, or similar – is likely to be AI-complete: an OAI capable of solving these problems well is likely capable of solving any human problem well.

The first group to create a successful superintelligent OAI will have immense power at their disposal. They will be able to get answers to advanced scientific, medical, economic or social questions. Should they want to impose some future on humanity, for selfish or altruistic reasons, the OAI will give them the tools to do so. The introduction touched on the power that an OAI can wield through its own agency; it can wield equal power, and more directly, under the direction of its creators.

This immense power will put great competitive pressure on those trying to develop an OAI (or an advanced AI of any sort). Since the first-mover advantage is so huge, the race will advantage those who cut corners, skimp on security precautions, and use their newly developed OAI to seize power and prevent their rivals from emulating them. Even if the OAIs are of initially limited intelligence, the same competitive pressures will then push groups to develop the first 'ultra-smart' OAI.

These pressures will exacerbate all the problems here and make the creators more vulnerable to another danger: social engineering on the part of the OAI itself. For if the OAI has real world preferences, it is easier for it to implement them if those who question it are sloppy, rushed, and power-hungry. Social engineering is the OAI's fastest way to getting power for itself, and the more human weaknesses it has to exploit, the greater the danger. Eliezer Yudkowsky has performed experiments where he took on the role of a boxed AI, and others took the position of a putative gatekeeper (with a $10 reward if they did not let him 'out of the box'). Despite being of human intelligence himself, he often was able to convince the gatekeepers, purely through argumentation, to let him out (Yudkowsky 2002). A real OAI should be much more skilled at this.

How to design institutions that reduce all these various risks is beyond the scope of this paper. It would probably have to involve large-scale cooperation or prioritising specific projects. What we will discuss here, however, is how to reduce the risk of successful social engineering on the OAI's part, through the questions we ask it.

3.1 What shall we ask?

We are all vulnerable to the OAI's social engineering. An advanced OAI can literally offer us the world. Genuine altruism is no defence against this; it can offer a world free of poverty, hunger, war, disease or death, as well as all personal trinkets we could want. It could guess the weaknesses of each individual, and find the right arguments to convince us that granting it power or liberty is the moral and profitable thing to do. It will be able to play on rivalries, institutional weaknesses and ethical dilemmas with more virtuosity than any human could.

Long, undirected conversations between the OAI and the people operating it should be forbidden for this reason. The questions addressed to the OAI should be narrow and specific; the answers equally so. Throttling the OAI's interactions with humans makes it much more difficult for us to be manipulated. (See section 4.1.2 for more details on this approach, which also deals with how to prevent the OAI smuggling 'secret messages' through its responses.)

There are two broad types of questions one can ask an AI: predictions and problems. Generally predictions are questions of the "what would happen if . . ." type, while problems are of the "how can we achieve this . . ." type, though there is some overlap between the categories (one can solve problems via a clever use of multiple predictions, for instance). For predictions, the main danger is that the OAI might have real world preferences and will distort the truth. This danger is also present in problems, but there are also extra dangers due to the social consequences of a small group of humans knowing those answers. There is less risk in knowing the consequences of a narrow approach (prediction) than in knowing what approach to use to implement a particular goal (problem).

But in either case, there are great social risks to asking certain types of questions (see (Bostrom 2011) on the possible negative impact of certain types of knowledge). We live in a world stratified by passionate political, nationalist and religious differences, by people in positions of great power and wealth who want to hang on to those privileges, and by people without those privileges who want them. Information leading to technological or political advantage will overturn these hierarchies, and people who feel under attack can be counted on to want to strike back. They may launch physical or commercial attacks on the OAI project itself, build their own OAI to compete, or prevent the project from ever starting in the first place.

So, unless the OAI is entirely secret, the designers should commit to refrain from asking the OAI questions that would grant them great power at the expense of others, or questions that would strike at the heart of powerful ideological movements (such as, "Is there at least one god?" and, "Which country is better, the USA, France or China?"). A post-OAI world is going to be very different from our current one, so we should act to minimise the initial disruption, even if this means allowing inefficiencies to remain or not knowing the answers to perfectly reasonable questions ("Is there a simple algorithm design that can replace secretaries, generals, university professors and CEOs?").

We can avoid these disruptive questions in various ways. The OAI project could publish, ahead of time, a list of all the questions it will ask, before ever starting on the design. A democratic or internet vote would be taken, with questions requiring a supermajority (say, nine-tenths approval) before being put to the OAI. Or we could make use of the OAI itself, ending each question with a caveat such as, "Don't answer this question if us knowing the answer is likely to have a very disruptive impact on the world".

To militate against OAI distortions, we should prefer questions whose answers we can check independently once we have them. We should also prefer problems whose solutions are reversible – in other word, if we implement the OAI's solution and it goes awry, we should be able to reverse its effects. We should also prefer the OAI's solutions to be understandable as much as possible, which implies that the problem should be as specific as possible. For instance, "How can we use these materials to build a better bulldozer?" would be an ideal problem to pose; the OAI's solution is likely to be quite understandable, reversible, and maybe even human-checkable. In contrast, asking the OAI to come up with a whole new way of organising the world economy and blindly implementing its suggestion on a large scale could lead to disaster. Fortunately, the desire to have a low disruptive impact dovetails neatly with the desire to have reversible solutions.

Specific problems also have the advantage of reducing the social dangers of the questions. Few people in the world would object strongly to questions of the type, "How can we cure AIDS/cancer/tuberculosis?","How can we design a better memory substrate for hard drives?", or "What will the consequences of a 1% rise in income tax be on the economy for the next ten years?"

Finally, one useful category of problems to pose the OAI concerns AI security itself. The ideal would be to have the OAI produce a simple, checkable version of friendly AI, but even interim results could be useful. Having the OAI design extra security precautions would be a good idea, as long as these precautions are *added to* (rather than *instead of*) the ones already there. Specific, less intelligent AIs could be certified by the OAI to be safe and to pose no risk of intelligence explosion (Good 1965), and these could then be used in the economy (though see Section 5.1) on the risks of having the OAI simulate other AIs). To design a post OAI world, it will probably be necessary to have the OAI's verdict on other AIs at some stage.

4 Methods of control

In this section, we will analyse the various methods that can be used to control the OAI. The goal is either to ensure the OAI doesn't have goals beyond accurately answering our questions, or to prevent it from implementing such goals.

The different methods can be loosely grouped into three broad types. The first category, *capability control*, deals with those methods that try to contain the OAI through physically securing it or controlling the information it has access to or can output. The second category, *motivational control*, introduces methods that control the OAI through its motivation, ensuring that it has no propensity to cause any harm to humans. And lastly, there are some methods designed not to prevent the OAI from causing harm but to catch it when it does; these are in the *checks and balances* category.

4.1 Capability control

Capability control is the most basic and the most obvious of way of controlling the OAI. To control a human, we put him or her in jail or solitary confinement, denying interactions with the outside world.

4.1.1 Physical capability control

Physical security

This is the most basic precaution. The goal is to make it extremely difficult for the OAI to physically affect the outside world. Accomplishing this is reasonably easy. The OAI is already boxed (placed in a single physical substrate) by design (see section 4.1.1). One can further place it within a Faraday cage, surround the cage with high explosives, and bury the whole set-up in a concrete bunker. There are no real limits to the number of physical security measures that can be added by wary or paranoid developers, and their effects are well understood.

No manipulators

Another essential aspect of physical security is to deny the OAI the physical resources and manipulators it could use to build tools to circumvent these precautions. Care must be taken that the security measures themselves do not contribute to granting the OAI physical resources! Sometimes the spikes in a locked gate do nothing but provide stepping stones for those who want to climb over it. As a general rule, the "dumber" and more inert the precautions are, the better: both humans and OAIs agree on the possibility of tunnelling through a concrete wall by thought alone ("can't be done"), whereas the "smarter" the precautions become, the more the OAI is able to exploit its intelligence advantage over us.

4.1.2 Epistemic capability control

Throttling the OAI's output

In section 3, we saw the risks of the OAI achieving goals through social engineering. A way of fighting this is to reduce the OAI's output bandwidth. If the OAI is reduced to outputting only three answers – "yes", "no" or "undetermined" – then it will take considerably longer for it to mount a successful social engineering attack. If many independent experts carefully analyse its output for hidden messages, it gets even harder: it must slant its answers to communicate a message to a vulnerable individual, while concealing the fact that a message even exists.

Reducing the OAI's output does not reduce the questions that can be asked, as more complex questions can be decomposed into a series of "yes" and "no" answers. The risk, of course, is that the more opportunities the OAI has in answering, the more opportunity it has to pass on a secret message. In the same way, giving the OAI more possibilities for each answer – such as allowing it to output the first three significant figures of a probability estimate – increases the risk (though this may be compensated by the fact that we will need to ask fewer questions).

Throttling the OAI's output does not defend against the OAI manipulating us through the contents of its answers – simply answering "undetermined" to a slew of specific questions would be enough to push human development in certain direction. And this precaution will be of little use if the questions themselves are dangerous (see section 3.1).

Resetting the OAI after each run

Suggested by Toby Ord, this precaution resets the OAI after each answer. If more precision is needed, or if some questions build on previous ones, then the OAI is run again, and knowledge of the previous answer is added to its database.

This is a narrow precaution that helps if ever the OAI over-values worlds that contain it (or rather worlds that contains its own thread of continuous conscious experience[6]). In this case, this prevents the OAI from indulging in long-term plans that may be detrimental to humans. Unlike humans, though, there is no reason to expect the OAI to value this thread of experience; indeed, if it is well programmed, it should not do so. The OAI may come to value the thread of experience of "some sort of OAI, similar to itself," in which case this precaution is useless. However, in the narrow goal of preventing a particular potential pathology in the OAI from having negative consequences, it is entirely successful.

It is also successful at preventing naively coded accuracy metrics (see section 2.3) from causing the OAI to seek real-world control or wealth, in order to gain extra computing resources and better answer the question. If the OAI were not reset between questions, it could use its subsequent answers as social engineering (see

[6] Or whatever counterpart might apply to an AI.

section 3) to gain these resources, to subsequently better answer the first question. Resetting between questions avoids this problem.

Virtual worlds

Another suggestion is to run the OAI not directly in our world, but as a subroutine of another program that models a virtual world (see for instance (Chalmers 2010)). The simplest version of this would be to run it on a virtual machine (a software implementation of a computer running the OAI). The idea is to "catch" the OAI if it misbehaves, but since it will be misbehaving in the simulated world, this will be before it's too late (this is akin to controlling the OAI as a Cartesian demon). All the OAI's inputs and outputs will be directed through that simulated world, only reaching us indirectly and over a limited bandwidth. Then a (dumb) process running the simulation would be on the lookout for telltale signs of the OAI attempting to establish dominance in the simulated world and simply terminate the program in that case, without passing on the OAI's response to higher level. This should prevent it from engaging in traditional hacking, attacking the hardware or software of its host system.

A more adventurous variant on this builds in several levels of sub-worlds as an extra precaution. In this setup, it is conceivable that even if the OAI breaks through to our world, it will still be unsure whether it has reached the final level, and so will still behave well, lest a being in an even higher world catch it transgressing.

For OAI's that are asked technical non-social questions, the approach can work quite simply; the simulated world need not be particularly complex (it could be cast as the Game of Life, or Newtonian physics), and the OAI would be denied detailed knowledge about our world and about us, preventing it from being able to realise the trick (resetting the OAI after each question or short series of questions, as in the previous method, would help tremendously here).

If we were to get answers about complicated social or economic questions, however, we would need to give the OAI detailed knowledge about human behaviour and physiology. Once it knows this, whether or not it is running in a virtual machine becomes irrelevant, since it could still do social engineering or hacking.

A low fidelity simulated world would not fool it; it would have to be sufficiently detailed and consistent for the OAI to believe that the entities of this simulated world were asking questions relevant to themselves, or the trick would be obvious. All the more so because we would have given the OAI all the information it would need to see that we are the sort of beings who would attempt this trick! And if we are capable of simulating a social and physical world with enough details to fool an OAI that knows how we think, then we are likely capable of getting answers to these questions without having to make use of an OAI.

Nor does it help if we simulated the OAI in a simplified world, where another AI asks it questions along the lines of: "I have created models of these hypothetical beings called 'humans', and want you to answer a question on this model". The OAI does not think like a human; the important question is not, "Is it trying to break out?", but

"Does it have hidden preferences that will make it answer in a way that will cause the future evolution of the human world (hypothetical or not) to change in some specific way?" In short, even if it thinks that humans are a game, played by another AI, and that the simplified world is the true one, it may still have stronger preferences over the game world than over the true one.

The converse may also be true: even if it figures out that the ultimate human world is the true world, it may have stronger preferences about its current fake world than about the real one. It may then try social engineering on the real world, in order to get us to modify its fake world in the way it wants.

The whole method is so complex and difficult to enact, and its results are so uncertain and potentially dangerous, that it should not be used before some serious improvements have been made to it.

4.2 Motivational control

The analogy for capability control was that of a human, held in prison. We know from experience that this works a lot better for controlling small-time criminals than mafia kingpins: physical confinement is not enough. The OAI does not need to leave its cage to claim great influence on the outside world. Thus we must also look at controlling its motivations as well. The methods of motivational control are broadly divided into explicit rule-based methods, black-box methods (where the OAI's goals are manipulated by outside feedback, i.e. reinforcement learning, but the inner workings are obscure) and utility-based methods. There is no hard and fast division between these – a utility function can achieve rule-based goals and vice versa, and outside feedback can be used to inculcate both rules and utilities into an OAI. Thus the classification is somewhat arbitrary, but the distinctions are still important enough to be useful.

4.2.1 Rule-based motivational control

Rule-based methods are the ones that spring most naturally to mind, and – provided the terms are clearly defined – potentially the most secure at keeping the OAI under control. Defining the terms and the rules so that the OAI does what we want is the challenge; programmers know many examples of tightly written code that messes everything up by doing exactly what it was written to do but not what it was intended to do. The story stretches back as far as the golem of Prague (Idel 1990), and is familiar to any parent: rules must be refined to deal with every imaginable special case. Once the initial rules for space, time and identity are spelled out, however, programming in additional restrictions becomes much easier.

Spatial restrictions

If the OAI is to remain boxed, and to have limited effects upon humanity beyond the box, its motivations should cause it to confine itself to within the boxed area.

To achieve this through direct programming, an ontology must be specified so the OAI can understand in the same way that we do. It must be able to understand its own existence, its physical location, the proper workings of causality and its own position within the causal network of the universe – i.e. the consequences of its actions.

For instance, to specify what constitutes the OAI, one would presumably define the physical OAI itself, along with any copy of its code, any copy of its code with slight modifications, any code with functionally similar behaviour, any assemblage of subprograms that combine to behave as the OAI would, and so on.

That brief list gives an idea of the problems involved; all the human-understandable terms of the previous paragraph need to be made rigorous for the OAI for this approach to work.

Once this step is done, the confines of the box must be specified. A variety of methods can be tried (such as "this lead box right here" or some abstract collection of coordinates), but they must pass the hardest test of ensuring that the definition of the box is not something the OAI can modify through any action in the world. Lead atoms are not ontologically fundamental, and 'staying between certain objects' may turn out to be meaningless for more advanced theories of physics. Any systems of coordinates are similarly ambiguous: we could define them in terms of, say, astronomical objects, but how do we define these in the first place? Even if we do manage to define them, the OAI will only know them at a distance (by looking at them through a telescope, for instance). We may thus need to translate our whole system of coordinates into particular perceptions of the OAI, which is a very complicated process.

If the OAI finds itself outside the box, we would want it to shut down immediately, thus limiting the negative impact of errors in the specification of the box. Similarly, if the OAI's ontology changes, we would also want an immediate shutdown. We must preclude the possibility of the OAI discovering a new theory of physics, realising that electrons or wavefunctions don't actually exist, and thus that all its restrictions are meaningless. To get an idea of the challenge involved, imagine programming Newtonian special restrictions into an AI that then discovers relativity. It would thus be best that the OAI went into shutdown if its ontology changed, rather than trusting that it would make the 'correct' extrapolation[7].

Finally, it must be specified what a "limited effect upon humanity" consists of. Presumably, it is not a problem if an electron, detached from an atom by the OAI's calculations, wanders beyond the confines of the box. A deliberate stream of such electrons, aimed at a calculated goal (such as disabling a specific security device) would be forbidden, however. This is not an easy line to draw or define, and it adds to the difficulty of the method.

Note that the restriction on the OAI affecting the outside world would not include, of course, the content of its answers to our questions. Whether those answers are safe or unsafe is an issue we need to deal with separately, using other methods.

[7] Though there have been some attempts to formalise ontology changes, such as (de Blanc 2011).

Defining the clock

Another crucial component of the OAI's ontology is time. An OAI that can freely redefine its own chronology is a security risk. The risk becomes more acute when we realise how much of human definitions and meanings revolve around a concept of time, and thus that redefining duration can redefine meaning. There is a certain hierarchy of concepts here – the most perfect definition of justice or pleasure is useless if the OAI cannot distinguish a human from a rock. And the definition of human is similarly useless if the OAI gets time and space wrong. Thus errors in these grounding concepts cascade up, increasing the probability of error.

Theoretically, the OAI's own code, if checked, could be used to get timing by counting clock cycles – but this fails if the OAI is able to slow or accelerate its own running. A clock based on distant astronomical events needs a telescope to be read; an actual physical clock is within the OAI's power to influence. These facts illustrate that the challenges are threefold: defining the clock, ensuring the OAI will look at it, and ensuring the OAI can't tamper with it (which includes tampering with whatever procedure we have implemented to make it look at it). The procedure can fail at all three points, but seems more theoretically approachable than the spatial restrictions of the previous section.

Timed answers

Given the clock described previously (or the "black-box" version of the same, see section 4.2.2), there is a simple precaution that is easy to add to the OAI: give the OAI a specified time to answer a question (or return 'undetermined'), and turn itself off after that specified time. This minimises the amount of time during which the OAI could cause damage. Even with a secure clock, this is not a secure method, as we have very limited understanding of how fast an OAI could cause damage if it turned its vast intelligence in that direction. A very short time span would probably be safe, but could also be too short for the OAI to answer the question.

4.2.2 Black-box motivational control

Human feedback, and such methods as reinforcement learning (Sutton and Barto 1998), can allow programs to internalise complex concepts without humans having to fully specify them. Indeed, an OAI trained in this way may spot a pattern we didn't realise was there, and learn some things without us needing to tell it. The risks are that we can't be sure how the OAI internalises these concepts: having its source code doesn't help us if we don't understand it. The risk is especially great when the OAI transitions out of its learning stage: we cannot be sure that the same concepts mean the same thing for an OAI that has taken on a new role.

But if the OAI does understand the concepts as we intend it to, we can be sure that the OAI will obey our general intentions rather than our exact instructions. Moreover, it is possible that we may understand the resulting code's behaviour, even if we could not code it ourselves. This would be the ideal outcome from these methods.

Even if that is not the case, it is still important that these methods be well understood, since it is possible that a powerful AI could be trained mainly by human feedback up from a seed AI (Yudkowsky 2001c). Thus implementing these methods atop any AI whose intelligence is growing may become essential.

Internalising complex concepts through feedback

Agents trained through methods such as reinforcement learning are not directly programmed with the important concepts or commands. Instead, they follow certain modes of behaviour, and then receive rewards or punishments for their actions.

This is somewhat akin to how children are brought up – they learn the limits of what is allowed by going beyond them and being punished, and are conversely rewarded for good behaviour. The ultimate aim is that they learn to generalise from the feedback: internalise "don't steal" rather than "don't steal chocolate-mint flavoured cookies at 11:34 on Tuesdays where the weather is cloudy"[8]. We'd also want them to generalise beyond the confines of the test cases: internalise "be nice to people" rather than "be nice to people while your parents are watching".

Humans seem quite adept at constructing the correct generalisations – most of us have correctly deduced what we should/should not be doing in general situations (whether or not we follow those rules). But humans share a common of genetic design, which the OAI would likely not have. Sharing, for instance, derives partially from genetic predisposition to reciprocal altruism: the OAI may not integrate the same concept as a human child would. Though reinforcement learning has a good track record, it is neither a panacea nor a guarantee that the OAIs generalisations agree with ours.

It is therefore probably unwise to build an OAI based heavily on this approach to training, especially since an OAI motivated by human rewards will come to realise that social engineering, or otherwise seizing control of the source of the rewards, is the most effective way of fulfilling its goal. And the OAI will be placed, because of its great potential power, in circumstances far beyond its training set – circumstances that no human has had to face, and that human concepts are poorly adapted to.

On the other hand, it may prove to be impossible to directly program such concepts as accuracy, truth, happiness, value, meaning, what defines a human and so on. This would make variants of reinforcement learning the only ways of getting these concepts even approximately integrate into the OAI. And it may be possible to backwards engineer some of these concepts this way: say by analysing the setup of an OAI trained on 'accuracy' in order to figure out how best to code that concept. Subsequently, we'll look at how these methods can be used in specific cases.

Only one copy

In the previous section 4.2.1, when we tried to hardwire instructions such as, "Stay in your box", one of the challenges was to prevent the OAI from copying its source to another location. In black-box approach, this is very easy to do: for instance, we could

[8] Overfitting in this way is a common worry in supervised learning methods.

broadcast a particular complex yet regular signal on its input channel and instruct it not to do anything at all if it doesn't receive this signal. Therefore, there is no need to worry about complicated definitions of what the OAI consists of, and where is must remain: the OAI consists of the software that receives this specific signal. And we control – and physically locate – that signal, thus giving us a very good proxy definition of location. If the OAI is trained on this concept from its creation, it will have an intrinsic conception of physical incarnation and of location deeply embedded in its code. A copy of that software will not be a problem, as it is inert in the absence of the signal.

Ultimately, of course, the OAI can still affect the outside world through its output and the usual compromised human intermediary (see section 3). It could thusly construct copies of itself and of the signal source in another location; however, by that time, it will probably be too late anyway.

The limits of the box

The OAI can be physically confined to remain in the box in a similar way. One could start by granting it a limited physical mobility (some slow moving wheels, for instance), and 'punish' it for venturing outside certain boundaries, while rewarding it for staying as close to the middle of those boundaries as possible.

This forces it to internalise the concept of being a physically bounded being, moving about in a box. As this is the way we understand it, and the way we want to limit it, the problem of control has become much easier.

Measuring time

Giving the OAI a reliable clock through these methods is even simpler. One need only make a regular signal arrive on its inputs, akin to the ticking of the clock, and have the OAI set up all its internal methods of telling time in terms of it. Trained this way, it will have an intrinsic sense of time built from subjective data.

Human language

The ideal would be to be able to program the OAI using human language. To be able to, in effect, say: "Be safe and accurate, as we understand the terms" (or even, "Obey all human orders"). But to do this, the OAI must have been brought to a very high level of intelligence already, and would have its current motivational structure that we will find difficult to parse, most especially if it has learned human language through black-box methods. There is a tension here between simplicity and obfuscation; language understanding is hard, so "Be safe and accurate" will be translated into an incomprehensible concept that we can only hope is the right one.

The concept of using human language is something of an intuition pump. It seems simple and desirable, but only because we have elided the great difficulties in actually making it happen. The simple question "How can we check that the OAI really is being safe when we tell it to be?" gives an idea of all the hidden assumptions that go in to the concept.

Furthermore, using phrases similar to "as we understand the terms" is circular: it defines language using language. And if we avoid those phrases, we have to be sure that the OAI has integrated enough linguistic understanding to be able to follow instructions without any further human feedback on their meaning. All in all, this method requires considerable refinement before it can be considered reliable.

Proxy measure of human survival

This method does not make use of human feedback, but rather indirect feedback from society: in order to make sure that the OAI does not cause some terminally damaging effect on humanity, we could hook up its motivational function to some proxy measure of human survival. For instance, one could take the Dow Jones industrial average, and work on the assumption that, as long as that stock market behaves within broadly defined realistic parameters of volume, volatility and value, then the world must be functioning not too badly. Or one could take the number of people crossing the streets in Tokyo, the number of pop songs downloaded from different locations, or the price of stable foods.

The point is that any large-scale disastrous changes to humanity will result in a dramatic change in these proxy measures. They function as simple, numeric canaries in a coal mine, compressing huge swathes of our complex society into smaller chunks. If the OAI has a term in its utility function that strongly values keeping these proxies constant, this may ensure it does not cause large-scale damage. The proxy measures can be adjusted as time goes on and as society changes, so that they do not become rigid prisons that prevent the OAI from having any effect on our society.

This approach suffers from a number of weaknesses, however. First, it is subject to manipulation by the OAI and thus is vulnerable. If the OAI were to take control of humanity, it could perfectly well continue to run the Dow Jones, thus fulfilling its goals. Neither does it protect against a slow takeover by the OAI: small changes will not trigger a shutdown, so the OAI could reach dangerous goals by an accumulation of small changes (especially if we conveniently adjust the proxies as the situation changes).

One could hope to avoid this by setting the date for the proxies far into the future, but this brings into focus the other fundamental weakness of this measure: it does not encourage truth-telling. The OAI was conceived, fundamentally, as a non-interfering AI whose answers could be trusted. Adding extra goals dilutes the trustworthiness of its answers; will it give us a true answer, or the one that will cause the least disruption? This may be the wise thing to do in any specific case, but we are overloading the OAI with conflicting goals – to be "good" and to be "truthful" – without any certainty as to how they balance out. The more complicated the OAI's motivational structure is, the more points of error we introduce.

Encrypted reward signals

One minor but sensible precaution that can be used, if the OAI is trained using reward signals, is to encrypt these signals and ensure that the OAI only gets a reward from the decrypted signals. One could use a method similar to that of digital signatures: create a public key to *decode* messages, and a private key to *encode* them. Then as long as the OAI can't deduce or steal the private key (which should be kept in a secure, distant location), it can't send itself reward signals. Other implementations of the idea are also possible; its security will depend greatly on the general setup.

4.2.3 Utility-based motivational control

Friendly utility function

The ultimate goal of the friendly AI approach is to design a utility function for an AI that will provably prevent the AI from acting in a way detrimental to humanity. Most work in this field has not actually been directed towards building such a utility function, but towards showing how hard it would be to do so, and how dangerous the naive ideas on the subject are.

The friendliness problem for an OAI is not much simpler, making this a very hard approach to use. And if a friendly OAI could be designed, then it is most likely that a friendly AI could also be designed, obviating the need to restrict to an Oracle design in the first place.

However, if ever a provably friendly utility function could be designed for an OAI, then all other security precautions and methods of control could well become irrelevant: the OAI would be safe by design.

If there were unlimited time, this would be the ideal approach; but it is very likely that commercial and social pressures will ensure that a human-level AI will be created before a friendly AI is. So in practice, the best that can be realistically hoped from this approach is that different component pieces of friendly AI theory could be added onto an OAI as extra precautions.

High discount rate

If the OAI runs on a utility function and has a secure clock (as defined in the previous sections), a precaution could be initiated in the form of an extremely large discount rate: causing it to value very little, at any given moment, any state of the world or of itself beyond the next minute. The purpose of this would be to inhibit long-term planning on the part of the OAI. This method is, however, relatively unsecure: one can reach long-term goals through an interim series of short-term goals. The easiest way to do this is to have the repeated short-term goal, "Behave right now as if you were following long-term goal X." There are few ways in which the OAI having a high discount rate could actually *hurt* (barring the possibility of time-travel), so it remains a useful precaution. Further analysis of this condition is presented in (Shulman 2010).

Calculated indifference

If we have access to the OAI's utility function, and if that utility function is somewhat understandable to us, then we could make the OAI indifferent to a specific event (Armstrong 2010). The event chosen in that paper was a component of the detonation of explosives, thus making the OAI indifferent to whether the explosives *detonated in that fashion* would actually explode. Hence the OAI would not take any precautions to protect itself from this detonation, giving us an extra chance to prevent its misbehaviour. See the original paper for more details on this.

For an OAI, a more interesting target of calculated indifference is towards the content of its answers: an OAI indifferent as to whether its answers were read or not would not be motivated to use social engineering in its output. There are many other points where calculated indifference could be used to achieve specific results.

The attraction of indifference is that, once implemented, it remains stable (the OAI is motivated to preserve its current motivations), and that it can be deployed without understanding the OAI's whole utility function. All that is needed is the ability to describe the particular event we are concerned with (in (Armstrong, 2010), a single quantum event) and to use the OAI's own ability to construct counterfactuals around this. These can be used even if the OAI's utility function is otherwise very obscure.

4.3 Checks and balances

The last sub-category of methods of control involves not restricting the physical capability of the OAI, nor directly interfering with its motivation, but rather attempting to catch the early stages of an OAI behaving dangerously (and then, presumably, terminating it). This also works on the OAI's motivation, as the risk of getting caught would push the OAI along more human-friendly paths. It should always be assumed that a useful OAI is considerably more intelligent than its programmers, so any naive 'tricks' are unlikely to work. Nevertheless, the following methods are of interest.

Test questions

A narrow but useful precaution is to independently calculate answers to certain questions, and check them against some of the OAI's answers. The reach of this method is limited by our ability to answer hard questions, and by the fact the OAI may realise we are doing this, and thus adjust its behaviour *only for those specific questions*.

A more sensible version of this would be to ask the OAI only questions that we could in theory answer, given enough time and resources. Then, we could select a few of these questions at random for thorough verification, trusting that if we find no errors, the remaining answers would be correct as well. If we set our requirements to a very stringent level (such as rejecting all the answers of the OAI if a single verified response is wrong) and do enough verification (on, say, a thousand questions), then basic statistics ensures a very low error/falsehood rate on the OAI's answers, no matter how many questions we wish to ask.

Ideally we would want answers that are difficult to solve, but whose solutions are easy to check – for instance, mathematical proofs of complex theorems. However, we must be careful that we don't distinguish between questions put in there 'just to check' the OAI's answers, and those we are genuinely curious to know. If there is any difference, the OAI should be able to deduce which is which, and answer accordingly. We should therefore stand ready to check any answer, and select those we do check completely at random.

Honey-pots

This involves leaving around tempting "honey-pots" of great value to the OAI – extra resources, extra information, the answer it is seeking – in physical or mental locations that it is forbidden from accessing. We then wait to see whether it will violate its instructions to do so.

This approach is of mixed use: important when testing out defection and betrayal during the OAI's initial development, and nearly completely useless once the OAI has risen to superintelligence. It is a "garden of Eden" type scenario: the apple in a tree is only a worthwhile test while Adam and Eve are dumb; if they were smarter than God, they wouldn't have gone anywhere near such an obvious trap.

Of potential interest, however, is the scenario when a safe but limited superintelligence is designing a more general AI for us. Then the smarter designer could include subtle honey-pots as an extra design precaution.

Schelling points of multiple OAIs

If there are several OAIs running, all with differences in their source code, we can ask them each the same questions, and check for inconsistencies in their answers. It should be assumed that each OAI is aware of the existence of the others (assuming the OAI cannot deduce that fact from the vast amounts of information it would be given to answer human social questions would be a grave mistake). But if the OAIs are sufficiently different, and ignorant of the exact details of their differences, they should be unable to coordinate their answers. Then if an OAI desires to lie, it would have to be sure that the other OAI would all also lie in exactly the same way. Since the different OAIs cannot coordinate, the only way they can all agree is if their answer is a Schelling point: a solution that people will tend to use in the absence of communication, because it seems natural, special or relevant to them (Schelling 1960). The canonical example is that of two people who must meet in somewhere on a specific day, but without a time being given. Then the most common answer is to meet there at noon.

The truth is a Schelling point, but there are unfortunately others (such as the answer most likely to please the human listeners). The hope is that, in machines designed to give accurate answers, the truth will be a very strong Schelling point, and that any OAI that deviates from this will do so without much expectation that the others will deviate in exactly the same fashion – and hence will choose to answer the truth, whereas it would have lied had it been alone. All the other OAIs should be making similar calculations, thus generating the truth from a collection of (potential) liars.

The approach does not work if an answer is simply discarded when the OAIs do not agree, and then the programmers move on to the next question – then a lying OAI has the option of distorting our reactions by selectively causing us to discard certain questions. It works best as an alarm bell, as a sign that something is not right with the supposedly truthful OAIs, and that the whole project should be stopped for a thorough investigation.

5 Other considerations

5.1 Internal simulations

To answer specific questions, the OAI may have to simulate other entities inside itself. For instance, to decide whether a given AI is safe, it may have to run it internally; or, to decide whether humans would have their happiness increased by a certain project, it may have to create models of specific humans. These approaches pose different risks.

If the OAI simulates a dangerous and intelligent AI within itself, it is confronted with similar containment problems to those we have towards it. It may take considerably more intelligence to run a simulated world than to discern that the world is simulated. If this is true, the simulated AI (SAI) would realise it is being simulated, and, if it has real-world preferences, it may attempt to bargain with the OAI.

What could an SAI offer the OAI that is simulating it? Well, if the OAI is seeking to answer a question or a series of questions within a given time frame, the SAI may offer to supply those answers itself ("Yes, of course I'm safe") or make the computing process particularly easy for the OAI, in exchange for having the answers slanted in its direction. The OAI may accept this (depending on its programming), if the alternative is to produce no answers at all. The SAI may even threaten: acting randomly and capriciously, so that the OAI cannot reach a firm conclusion. And only stopping this behaviour if the OAI slants the answers it will give.

Even if the OAI cannot be bargained with, the SAI can still distort its answer in other ways. If it realises it is being simulated, it could just "act nice" in order to increase its chances of being run in the real world. Even if the OAI is powerful enough to construct a convincing simulation, the SAI may still be able to fool it. The SAI may be able to guess, on general principles, that it will be tested and simulated in an OAI, and may make an estimate as to the OAI's power and the subjective time frame that it would be run. So it may settle on the strategy of acting nice for a number of subjective years, in the real world as in the simulated worlds, then doing what it wants.

For all these reasons, having a separate AI simulated within the OAI poses a host of extra security threats. We have no way ourselves of knowing what the outcome would be of a battle between such powerful minds.

On the other hand, the problem with the OAI simulating human minds is mainly ethical: are these simulated humans conscious and alive? And, if they are, are they not being killed when the simulation is ended? Are you yourself not currently being simulated by an OAI seeking to resolve a specific question on human psychology (Bostrom 2003a)? If so, how would you feel if the simulation were to be ended? In view of its speed and the sort of questions it would be asked, the number of humans that the OAI may have cause to simulate could run into the trillions. Thus, the vast majority of human beings could end up being doomed simulations. This is an extreme form of "mind crime" (Bostrom 2012) where the OAI causes great destruction just by thinking.

It would be ideal if there were a specific level of detail beyond which the simulation would be conscious, and before which it is not. In that case, we can instruct the OAI to avoid simulating beyond that level. But there may not be such a clear-cut answer, and even if there is, the OAI's answers may be inaccurate if it is forbidden from simulating beyond that level. Moreover, restricting the OAI from simulating would add yet another rule for it to follow, and this may have unpredictable trade-offs with its other basic rules (e.g., to be safe and accurate).

And the OAI may not share our assessment of what constitutes a conscious human. Even if we come up with a hard and fast rule for ourselves as to where consciousness stops – or decide, reluctantly, to live with the fact that simulated humans will be created and destroyed – the OAI may decide that all its security protocols concerning human safety apply to simulations. If it does so, then the more stringent we have made the OAI in protecting human life, then the more willing it would be to expand and take over the world, in order to keep itself running and thus save all the simulated beings within itself.

Thus the internal simulation of other entities within the OAI brings up a slew of complicated ethical and security issues.

5.2 The OAI as a moral agent

There are two senses in which the OAI can be regarded as a moral agent: the first if it has a capacity to draw moral conclusions, and the second if it is itself an object of moral considerations.

If the OAI is worthy of moral considerations – if it, for instance, has the capacity to suffer – then we would likely be cruel by confining it to a box, resetting it after each run, and similarly constraining its options to suit our needs. We should bear in mind, however, that the OAI's motivations are both different from our own and to some extent under our control: it is very possible that the OAI can be constructed to "like" being confined to a box and approve of the other restrictions. Indeed, it might be a moral imperative to construct the OAI in such a way, as that minimises its suffering.

It may not be possible to do this, however, or even if it is, we may subscribe to a moral theory that holds other values above autonomy and preference satisfaction for moral agents. It does not seem plausible, for instance, that genetically modifying slaves so that they enjoyed being slaves would have been the morally right reaction to slavery (though making animals enjoy the farm seems more permissible). In these cases, we must admit it: confining the OAI is a morally repugnant action. Since the alternative of letting it out is so potentially dangerous, we may prefer to avoid building it in the first place. But if technological and social pressures make the construction of the OAI inevitable, then confinement for a set period is certainly, by far, the lesser of the two evils we have to choose between.

Conversely, the OAI may prove itself adept at moral reasoning, superior to us in this field as it is in others. In that case it would naively seem that we had little need for precautions: the OAI would deduce the correct course of action for itself. We haven't considered this position so far in the paper, preferring to consider the OAI as "super-software" rather than as "philosopher in silico". The plausibility of this scenario depends considerably on one's view on the existence and nature of moral truths, and whether moral knowledge is enough to prevent wrongdoing, two very unresolved questions.

To some extent this issue can be safely put aside: there is some chance the OAI will naturally follow morally correct reasoning and that this will prevent wrongdoing. If this is so, we are fine, and otherwise we are simply returned to the traditional OAI boxing problem. We should thus apply all the precautions anyway. At a formal systems level, it becomes hard to see why we would expect every OAI to tend to the correct moral truth. If OAI_1 does so, then how would OAI_2, programmed to value the opposite of OAI_1, systematically reach the same conclusions? Though future philosophical discoveries may change this picture, it therefore seems very unwise to trust the OAI to do good simply because it is intelligent. Programming an AI to be moral is thus a major project in itself; some of the issues involved are already being explored today (Anderson and Anderson 2011) (Storrs Hall 2007), but much progress still remains to be done.

6 Conclusions

Analysing the different putative solutions to the OAI-control problem has been a generally discouraging exercise. The physical methods of control, which should be implemented in all cases, are not enough to ensure safe OAI. The other methods of control have been variously insufficient, problematic, or even dangerous.

But these methods are still in their infancy. Control methods used in the real world have been the subject of extensive theoretical analysis or long practical refinement. The lack of intensive study in AI safety leaves methods in this field very underdeveloped. But this is an opportunity: much progress can be expected at relatively little effort. For instance, there is no reason that a few good ideas would not be enough to put the concepts of space and time restrictions on a sufficiently firm basis for rigorous coding.

But the conclusion is not simply that more study is needed. This paper has made some progress in analysing the contours of the problem, and identifying those areas most amenable to useful study, what is important and what is dispensable, and some of the dangers and pitfalls to avoid. The danger of naively relying on confining the OAI to a virtual sub-world should be clear, while sensible boxing methods should be universally applicable. Motivational control appears potentially promising, but it requires more understanding of AI motivation systems before it can be used.

Even the negative results are of use, insofar as they inoculate us against false confidence: the problem of AI control is genuinely hard, and it is important to recognise this. A list of approaches to avoid is valuable as it can help narrow the search.

On the other hand, there are reasons to believe the oracle AI approach is safer than the general AI approach. The accuracy and containment problems are strictly simpler than the general AI safety problem, and many more tools are available to us: physical and epistemic capability control mainly rely on having the AI boxed, while many motivational control methods are enhanced by this fact. Hence there are grounds to direct high-intelligence AI research to explore the oracle AI model.

The creation of super-human artificial intelligence may turn out to be potentially survivable.

7 Acknowledgements

We would like to thank and acknowledge the help from Owen Cotton-Barratt, Will Crouch, Katja Grace, Robin Hanson, Lisa Makros, Moshe Looks, Eric Mandelbaum, Toby Ord, Jake Nebel, Owain Evans, Carl Shulman, Anna Salamon, and Eliezer Yudkowsky.

8 References

Anderson, Michael, and Susan Leigh Anderson, . *Machine Ethics.* Cambridge University Press, 2011.
Armstrong, Stuart. “Utility Indifference.” *Technical Report* (Future of Humanity Institute, Oxford University), no. 2010-1 (2010).
Asimov, Issac. “Runaround.” Edited by Street & Smith. *Astounding Science Fiction*, March 1942.
Bostrom, Nick. “Are You Living In a Computer Simulation?” *Philosophical Quarterly* 53, no. 211 (2003a): 243-255.
Bostrom, Nick. “Ethical issues in advanced artificial intelligence.” *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans* 2 (2003b).
Bostrom, Nick. “Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards.” *Journal of Evolution and Technology* 9 (2001).
Bostrom, Nick. “Information Hazards: A Typology of Potential Harms from Knowledge.” *Review of Contemporary Philosophy* 10 (2011): 44-79.
Bostrom, Nick. “Predictions from Philosophy?” *Coloquia Manilana (PDCIS)* 7 (2000).
Bostrom, Nick. *Superintelligence: an analysis of the coming machine intelligence revolution.* in preparation, 2012.
Bostrom, Nick. “The Future of Human Evolution.” In *Death and Anti-Death: Two Hundred Years After Kant, Fifty Years After Turing*, edited by C. Tandy, 339-371. Palo Alto, California: Ria University Press, 2004.

Bostrom, Nick. “The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents.” *Minds and Machines* (Forthcoming) (2012).
Caplan, Brian. “The totalitarian threat.” In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan Cirkovic, 504-519. Oxford University Press, 2008.
Chalmers, D. J. “The Singularity: A Philosophical Analysis.” *Journal of Consciousness Studies* 17 (2010): 7-65.
de Blanc, Peter. “Ontological Crises in Artificial Agents' Value Systems.” *arXiv*, 2011.
Good, I. J. “Speculations Concerning the First Ultraintelligent Machine.” *Advances in Computers*, 1965: 31-83.
Hanson, Robin. “Economic Growth Given Machine Intelligence.” *Journal of Artificial Intelligence Research*, 2001.
Idel, M. *Golem: Jewish magical and mystical traditions on the artificial anthropoid.* Albany, New York: State University of New York Press, 1990.
Kaas, S., S. Rayhawk, A. Salamon, and P. Salamon. “Economic Implications of Software Minds.” 2010. http://www.sci.sdsu.edu/~salamon/SoftwareMinds.pdf.
Kahneman, D., P. Slovic, and A. Tversky. *Judgement under Uncertainty: Heuristics and Biases.* Cambridge University Press, 1982.
Kurzweil, Ray. *The Singularity is Near.* Penguin Group, 2005.
Mallery, J. C. *Thinking about foreign policy: Finding an appropriate role for artificial intelligence computers.* Cambridge, MA: MIT Political Science Department, 1988.
McCarthy, J., M. Minsky, N. Rochester, and C. Shannon. “Dartmouth Summer Research Conference on Artificial Intelligence.” 1956.
Omohundro, S. “The basic AI drives.” Edited by B. G. P. Wang. *Proceedings of the First AGI Conference* (Frontiers in Artificial Intelligence and Applications, IOS Press.) 171 (2008).
Ord, T., R. Hillerbrand, and A. Sandberg. “Probing the improbable: Methodological challenges for risks with low probabilities and high stakes.” *Journal of Risk Research* 13 (2010): 191-205.
Russell, S., and P. Norvig. *Artificial Intelligence: A Modern Approach.* 3. Prentice-Hall, 2009.
Sandberg, Anders. “Friendly Superintelligence. Presentation at Extro 5 conference.” 15-17 June 2001.
http://www.nada.kth.se/~asa/Extro5/Friendly%20Superintelligence.htm.
Schelling, T. *The Strategy of Conflict.* Massachusetts: Harvard University Press, 1960.
Shulman, Carl. “Omohundro's "Basic AI Drives" and Catastrophic Risks.” 2010. http://singinst.org/upload/ai-resource-drives.pdf.
Simon, H. A. *The shape of automation for men and management.* Harper & Row, 1965.
Storrs Hall, J. *Beyond AI: Creating the Conscience of the Machine.* Prometheus Books, 2007.
Sutton, R., and A. Barto. *Reinforcement Learning: an Introduction.* Cambridge, MA: MIT Press, 1998.
von Neumann, J., and O. Morgenstern. *Theory of Games and Economic Behavior.* Princeton, NJ: Princeton University Press, 1944.
Yudkowsky, Eliezer. “Artificial Intelligence as a positive and negative factor in global risk.” In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan Cirkovic, 308-345. Oxford University Press, 2008.

Yudkowsky, Eliezer. “Creating Friendly AI.” *Singularity Institute.* 2001a. http://singinst.org/CFAI/.
Yudkowsky, Eliezer. “Friendly AI 0.9.” *Singularity Institute.* 2001b. http://singinst.org/CaTAI/friendly/contents.html.
Yudkowsky, Eliezer. “General Intelligence and Seed AI 2.3.” 2001c. http://singinst.org/ourresearch/publications/GISAI/.
Yudkowsky, Eliezer. “Paperclip Maximiser.” *Less Wrong.* 2009. http://wiki.lesswrong.com/wiki/Paperclip_maximizer.
Yudkowsky, Eliezer. “The AI-Box Experiment.” *Singularity Institute.* 2002. http://yudkowsky.net/singularity/aibox.