Risks and Mitigation Strategies for Oracle AI

Stuart Armstrong

Abstract. There is no strong reason to believe 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 goals or motivation systems. Oracle AIs (OAI), confined AIs that can only answer questions, are one particular approach to this problem. However even Oracles are not particularly safe: humans are still vulnerable to traps, social engineering, or simply becoming dependent on the OAI. But OAIs are still strictly safer than general AIs, and there are many extra layers of precautions we can add on top of these. This paper looks at some of them and analyses their strengths and weaknesses.

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

1 Introduction

While most considerations about the mechanisation of labour has 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) which do not exist for an artificial system. Beside 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 modelling), or use abilities not feasible to 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 N. , The Future of Human Evolution, 2004).

V.C. Müller (Ed.): Philosophy and Theory of Artificial Intelligence, SAPERE 5, pp. 335–347.
springerlink.com

The likelihood of both superintelligent and human level AI are hotly debated – it isn't even clear if the term 'human level intelligence' is meaningful for an AI, as its mind may be completely alien to us. This paper will not take any position on the likelihood of these intelligences, but merely assume that they have not been shown to be impossible, and hence that the worrying policy questions surrounding them are worthy of study. Similarly, the paper will not look in detail at the various theoretical and methodological approaches to building the AI. These are certainly relevant to how the AI will develop, and to what methods of control will be used. But it is very hard to predict, even in the broadest sense, which current or future approaches would succeed in constructing a general AI. Hence the paper will be looking at broad problems and methods that apply to many different AI designs, similarly to the approach in (Omohundro, 2008).

Now, since intelligence implies the ability to achieve goals, we should expect superintelligent systems to be significantly better at achieving their goals than humans. 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 N. , Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards, 2001).

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. These are not strategies that can be heavily relied on, obviously. Other approaches, such as balancing superagents or institutions (Sandberg, 2001) or "friendly utility functions" (Yudkowsky E. , Creating Friendly AI, 2001) (Yudkowsky E. , Friendly AI 0.9., 2001), are underdeveloped.

Another solution that is often proposed is the so-called Oracle AI (OAI)[1]. 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 AI that acts according to their own goals are obviously dangerous, oracles appear more benign[2]. While owners could potentially use them in selfish or destructive ways – and their answers might in themselves be dangerous (Bostrom N. , 2009) – they do not themselves pose a risk. Or do they?

This paper attempts to analyse the problem of "boxing" a potentially unfriendly superintelligence. The key question is: how dangerous is an Oracle AI, does boxing help, and what can we do to reduce the risk?

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

[2] Some proponents of embodiments might argue that non-embodied AIs are impossible – but it is perfectly possible to have an embodied AI limited to a particular "box" and have it only able to interact with the world outside the box through questions.

2 Power of AI

A human-comparable mind instantiated in a computer has great advantages over our biological brains. For a start, it benefits from every iteration of Moore's Law, becoming faster at an exponential rate as hardware improves. I.J. Good suggested that the AI would become able to improve their own design, thus becoming more intelligent, leading to further improvements in their design and a recursive intelligence explosion (Good, 1965). Without going that far – it is by no means obvious that the researcher's speed of thought is the dominating factor in Moore's Law – being able to reason, say, a thousand times faster than a human would provide great advantages. Research of all types would become much faster, and social skills would be boosted considerably by having the time to carefully reason out the correct response. Similarly, the AI could have access to vast amounts of data, with huge amounts of information stored on expanding hard drives (which follow their own Moore's Law (Walter, 2005)). So an AI would be able to think through every response thoroughly, carefully researching all relevant data, without any humans-noticeable slow-down.

Software can not only be run faster with more data, it can also be copied and networked. An AI need be only trained in a particular skill once; from that point on, it can be copied as much as required. Similarly, if AIs are subject to human-like vicissitudes, such as fatigue or drop in motivation, this can be overcome by taking the entity at the peak of its energy or motivation, and reloading this every time the AI starts to weaken. One could use this, for instance, to select AIs at moments when they are particularly successful at group-interactions. Thus a group of AIs, trained in different skills and with compatible motivations, could be networked together into a super-committee. Such super-committees are a likely stable organism (see (Shulman C. , 2010) for the full version of this argument) and could become what people refer to as "an AI". Their abilities would likely be a superset of any human committee: a major bottleneck in human organisations is the ability to share information rapidly and work well together. Thus no human group could out-think them, and with the ability to be trained once and copied at will, they could be used in any and all roles that humans fill today.

Another factor to consider is that evolution has made humans very skilled at social abilities and spatial recognition, but less so at scientific and technical abilities. So a "human comparable AI" that matches our social skills (a la Turing test (Turing, 1950)) is likely to be much more skilled than us at these scientific and technical tasks; even if it isn't, it's likely to have a superior ability to interact with other softwares that do have these skills.

Above-human intelligence AIs would be even more impressive, of course, if ever they existed. If we just constrained the AI into a box, interacting with it only through text messages, then we would be safe, surely?

3 Introducing the Oracle

That is precisely what an Oracle AI is: a confined AI with no physical manipulators, with which we can only interact through text messages. Though it feels superficially safer, this still leaves us vulnerable to the OAIs most dangerous tool: social engineering, the OAIs fastest way of getting power for itself.

Even humans can get themselves 'unboxed' purely through arguing. Eliezer Yudkowsky has performed experiments where he took on the role of a boxed AI, and others took the position of a putative gate-keeper (with a $10 reward if they did not let him 'out of the box', meaning granted it full access to the outside world). Despite being of human intelligence himself, in three out five attempts, he was able to convince the gatekeepers, purely through argumentation, to let him out (Yudkowsky E. , 2002).

A real OAI should be much more skilled at this: we should think of the OAI as composed of a committee of extremely smart and convincing people, working together perfectly at high speeds. Or, to shift the intuition pump, if we were the OAI, with these sorts of intellectual resources, would we not be able to persuade, trick or trap ourselves out of the box? Most people can think of three or four plausible methods that could work; we can rest assured the OAI would think of these, and better ones, and be capable of implementing it to the best of its ability, if it wanted to.

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.

On top of that, the OAI can always trick us with brilliant technology with undetectable 'poison pills', or simply become so important to human economy or human society that we dare not turn it off, whatever it does. In no way could a superintelligent AI be considered 'safe' merely because it is boxed.

4 Why Box the Oracle At All?

In view of the preceding, one could wonder why one should study Oracles at all. If the Oracle will inevitably convince or trick its way out of the box if it wants to, why bother investing in this type of precaution – shouldn't all our efforts be devoted to preventing general AI from being developed in the first place? It is unlikely that we can convince the world to back off from AI research, or permit the draconian checks that would reliably prevent narrow AI research from leading to general artificial intelligence. But trying along these lines is likely to have a tiny but real effect, whereas worrying more about Oracles seem to have no impact.

This is too extreme a position to take. Precautions are useful, even if they are imperfect. A well designed physical boxing system would prevent the OAI from escaping other than through social engineering. And there are myriads of clever ways an OAI could physically escape, if it had access to a manufacturing base; taking these off the table is a definite achievement.

Secondly, a safe Oracle is strictly easier that a safe autonomous AI, even one that also lacked a manufacturing capability. To see this, consider all the different groups in the world with different agendas and preferences, from libertarians to communists, democrats, nationalists, Kantians, utilitarians, religious groups, neo-Nazis, environmental groups, advocates of free love... The list goes on. If any of these groups were to build an autonomous AI that they felt was "safe", it would have to be imprinted with most of their values, which would differ from the values of the other groups. But if they wanted to build a safe OAI, they would all agree on what this constituted: an AI that stayed confined in a box and answered questions accurately. To phrase this colloquially, the difference between Hitler and Gandhi is not relevant for safe OAI (but is relevant for safe AI). It is difficult to estimate how much easier an OAI is, but if we use approximate Kolmogorov complexity (Solomonoff, 1960) as a heuristic, then we can see that the values and desires of these different groups require a lot of description, making general safe AI many orders of magnitude harder than safe OAI. We don't need to take this argument too literally in order to agree with the general point: pushing for OAI over general AI can make a huge difference.

But lastly, we don't need to just box the OAI, pick the least corruptible gatekeeper we can find, and hope for the best. Once boxed, there are many supplementary methods of control we can add on. Methods of capability and motivational control can all be layered over the basic AI framework, and we can add checks and traps to ensure that the OAI is actually behaving well. Looking at these "methods of control" will be the subject of the next section, though a more thorough analysis can be found in our paper (Armstrong, Sandberg, & Bostom, Thinking Inside the Box: Using and controlling an Oracle AI, 2011), which elaborates on all these methods.

5 Methods of Control

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 of control the OAI through its motivation, ensuring that it has not 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. It should be noted that, depending on the design of the OAI, some methods may be impossible (such as rule or utility based motivational control). Since the design of the OAI is not known, we won't rule out any methods on these grounds.

5.1 Capability Control

Capability control is the most basic and the most obvious of way of controlling the OAI. To control a human, we put them in solitary confinement, denying them interactions with the outside world. Similarly for an OAI, the first, minimum steps should be to restrict its means of influencing the world: that is essentially what boxing *means*.

The most basic methods are physical capability control. The OAI should be confined to a box. The 'box' might be a Faraday cage, filled with inert gases (or with nothing at all). The OAI should be denied manipulators, and ideally would have no/little control over any apparatus with moving parts – anything to prevent it from building a manufacturing capability. The cage could be surrounded with high explosives, triggered by some willing gatekeeper, and the whole set-up could be buried 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 – though we are far more likely to succeed with 'dumb' precautions: an OAI could conceivably hack an extremely advanced security lock with a blinking light, but couldn't kick down a wall without limbs.

Another possibility is to use epistemic capability control, to reduce the OAI's epistemic rather than physical abilities. It would be very useful to throttle the OAI's output, reducing the bandwidth it can use. 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. The risk, of course, is that the more opportunities the OAI has in answering, the more opportunity there is of it passing 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 less follow-up questions). And it does not prevent the OAI from manipulating us through the contents of its answers, pushing human development is certain directions.

Other epistemically plausible measures, in order of decreasing usefulness (Armstrong, Sandberg, & Bostom, Thinking Inside the Box: Using and controlling an Oracle AI, 2011) would be to reset the OAI after each run, add 'fake levels of reality' by running on virtual machines inside other virtual machines, and throttling the OAI's input. Resetting the OAI prevents *certain types* of long term plans from emerging while fake levels of reality are useful if the OAI is a highly technical entity asked highly technical questions.

Throttling the OAIs input is likely to be less useful: the OAI can deduce huge amounts of information from any question we ask and any information we give it. This can be best seen from the Kolmogorov complexity perspective (Solomonoff, 1960)). Since we cannot compute the Kolmogorov prior, but any superintelligence would approximate it better than we can, we cannot know how much information we are inadvertently giving it through our inputs (outside of certain closed situations). The potential for considerable undetected giveaways is immense, and these are errors we couldn't even know we had committed.

5.2 Checks and Balances

This sub-category of methods of control involve not restricting the physical capability of the OAI, nor directly interfering with its motivation, but attempting to catch the early stages of an OAI behaving dangerously (and then, presumably, terminate 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, we can get some mileage by testing a random selection of the OAI's answers. If we only gave the OAI questions that we ourselves could solve, given enough time, and if we exhaustively tested, say, a thousand of the answers it gave, then we would have a pretty good estimate of its trustworthiness, no matter how many questions we asked. The most important aspect of this is that the tested questions be selected at random, after the OAI has given its answers (a sort of 'double blindness' for questions). Asking specific test questions is a no-no; if the OAI can deduce that a certain question will be tested, it may answer differently on it.

Less convincing is the idea of leaving 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. And then waiting 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 super-intelligence. 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.

Finally, we might play multiple (but slightly different) OAIs against each other, by asking them the same question while preventing them from communicating with each other. The OAIs are then confronted with the choice between answering truthfully and telling a lie; but for the lie to work, all OAIs would have to stumble upon the same one. If the OAIs are quite different, and especially if their motivations are different and unknown to each other (a rare case where throttling the OAIs input will be of use), then they are unlikely to stumble upon the same lie, and may be compelled to tell the truth.

5.3 Motivational Control

The main area for potential improvement in OAI control rests in motivational control. Capability control is useful, but we probably already have a good idea of how to do it. Checks and balances are probably too uncertain to rely on. So the gains from further study in these areas are probably minor.

If we could solve the motivational control problem completely, then we would solve the OAI problem completely: we would have a friendly AI (Yudkowsky E. , 2001a) (Yudkowsky E. , 2001b). But even partial improvements can increase our chances of a safe OAI, especially given the other methods of control we have.

It is essentially a problem of reverse ethics: we know approximatly what we want the OAI to do (stay inside the box, answer questions truthfully, don't plot against us). And we seek to find the OAI motivations that will result in this happening. Then, of course, once we have found the correct motivations, we want to ensure that the OAI actually follows them.

There is a certain tension between these two goals: motivations which will most likely be the safest are not those which are easy to check that the OAI is implementing. The grading runs from black box motivations (such as reinforcement learning), through rule-based motivations (such as direct programming) to utility function based motivations. It is much easier to "teach" an AI to be safe through reinforcement learning than to spell out explicitly the utility function that would accomplish the same thing. But conversely, it is much easier to check that an AI is implementing a utility function than to be sure that the messy incomprehensible result of reinforcement learning is actually doing what we hope it is.

5.3.1 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.

The biggest challenge is to define space and time. If the OAI can have a correct understanding of space and time, many other restrictions can be easily added to it. But to achieve this through direct programming, an ontology must be specified, that 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. There is a similar challenge with 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. To illustrate with an admittedly trivial example, if the OAI had the reasonable-sounding moral premise that "painlessly killing a human being, who is going to die in a micro-second anyway, in order to gain some other good, is not a crime", we would not want it to be able to redefine millennia as seconds. Nor would we

want it to do this when considering the length of happiness or pain that it grants or imposes on us. More subtle errors will certainly creep up in more complicated scenarios; we have had little experience dealing with beings that can redefine their own clocks, so we have not had to spell out how many of our assumptions rest on this not happening.

Once time and space are defined, other problems become easier – though not easy. Confining the OAI to stay in the box involves defining "the OAI" (presumably including all exact copies, all similar copies, all instantiation of algorithmic systems that would be similar to the OAI, and maybe those that would have similar goals) and "the box" (an easier problem) and "staying in" (a harder problem – waste heat reaching the outside world from the OAI is ok, but building a laser to blast the outside is not). We also must preclude the possibility of the OAI discovering a new theory of physics, realising that 'electrons' or 'wave-functions' don't actually exist, and thus that all its restrictions are meaningless.

With space and time well in hand, we can define other rules, requiring the OAI to answer within a certain time frame, to use only specific resources (if we can define those), and to respect other constraints we may want to define.

5.3.2 Black-Box Motivational Control

Human feedback, and such methods as reinforcement learning (Sutton & 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 'know some things without us needing to tell it'. The risks are that we can't be sure how the OAI internalises these concepts: knowing its 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 the spirit and not the letter of these restrictions. Moreover, it is possible that we may understand the resulting code's behaviour, even if we could not code it in ourselves: this would be the ideal outcome from these methods.

Even without that unlikely ideal, it is important that these methods be well understood, since it is very possible that a seed AI (Yudkowsky E. , General Intelligence and Seed AI 2.3) will be trained mainly by human feedback.

Some of the problems touched upon for rule-based motivations are easier to solve using black-box methods. The AI can be trained to recognise time through the tick of a clock; to internalise spatial restrictions through experience in embodiment (Steels & Brooks, 1995) or to come to think of themselves as "one copy" through similar methods. These methods are somewhat more robust to changes in ontology, not being so dependent on getting the definitions exactly right.

There still remain grave risks for complex questions whose definitions are unclear to us, and when the OAI is placed in situations far removed from its training environment. And this approach also trains the OAI in the skills and

usefulness of social engineering: it will learn that certain specific responses increase the likelihood of reward, and will extensively model the complex system – us – that is most responsible for these rewards.

5.4 Utility-Based Motivational Control

The ultimate goal of the friendly AI project (Yudkowsky E. , 2001a) (Yudkowsky E. , 2001b) 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. In practice, however, the best that can be realistically hoped for is that different component pieces of friendly AI theory could be added onto an OAI as extra precautions.

Even without a full friendly utility function, there are some tools specifically available for utility based OAI. For instance, we could impose a high discount rate on the utility, to prevent the OAI from planning to far ahead. Other more exotic manipulations of the OAIs utility function should also be possible; one is detailed in (Armstrong, Utility Indifference, 2010). This updates the OAI's utility function so that it is indifferent to the consequences of a single probabilistic event, and could be used to make it indifferent to the explosives packed around it, or, more interestingly for an Oracle, indifferent to the consequences upon humans of its answers.

5.5 Extra Measures

An important aspect of the whole OAI design is that there are many extra methods that can be implemented and added on top of the basic measures. Exotic methods such as proxy measures of human survival and utility indifference are detailed in our paper (Armstrong, Sandberg, & Bostom, Thinking Inside the Box: Using and controlling an Oracle AI, 2011).

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 downright dangerous.

It is not a question of little hope, however, but of little current progress. 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. There is no reason that a few good ideas would not be enough to put the concepts of space and time on a sufficiently firm basis for rigorous coding, for instance.

And even the seeming failures are of use, it they have inoculated us against dismissive optimism: the problem of AI control is genuinely hard, and nothing can be gained by not realising this essential truth. A list of approaches to avoid is invaluable, and may act as a brake on AI research if it wanders into dangerous directions.

On the other hand, there are strong 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 strong grounds to direct high-intelligence AI research towards the oracle AI model.

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

Acknowledgements. I would like to thank and acknowledge the help of Anders Sandberg, Nick Bostrom, Vincent Müller, Owen Cotton-Barratt, Will Crouch, Katja Grace, Robin Hanson, Lisa Makros, Moshe Looks, Eric Mandelbaum, Toby Ord, Carl Shulman, Anna Salomon, and Eliezer Yudkowsky.

References

Armstrong, S.: Utility Indifference. FHI Technical Report (2010)

Armstrong, S., Sandberg, A., Bostom, N.: Thinking Inside the Box: Using and controlling an Oracle AI (2011); accepted by Minds and Machines

Asimov, I.: Runaround. Astounding Science Fiction (1942)

Bostrom, N.: Ethical issues in advanced artificial intelligence. Cognitive, Emotive and Ethical Aspects of Decision Making in Humans 2 (2003)

Bostrom, N.: Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards. Journal of Evolution and Technology 9 (2001)

Bostrom, N.: Information Hazards: A Typology of Potential Harms from Knowledge (2009), http://www.nickbostrom.com/information-hazards.pdf

Bostrom, N.: Predictions from Philosophy? Coloquia Manilana (PDCIS) 7 (2000)

Bostrom, N.: The Future of Human Evolution. In: Tandy, C. (ed.) Death and Anti-Death: Two Hundred Years After Kant, Fifty Years After Turing, pp. 339–371. Ria University Press, California (2004)

Bostrom, N., Salamon, A.: The Intelligence Explosion. Retrieved from The Singularity Hypothesis (2011), http://singularityhypothesis.blogspot.com/2011/01/intelligence-explosion-extended.html

Caplan, B.: The totalitarian threat. In: Bostrom, N., Cirkovic, M. (eds.) Global Catastrophic Risks, pp. 504–519. Oxford University Press (2008)

Chalmers, D.J.: The Singularity: A Philosophical Analysis (2010), http://consc.net/papers/singularity.pdf

Cook, S.: The complexity of theorem proving procedures. In: Proceedings of the Third Annual ACM Symposium on Theory of Computing, pp. 151–158 (1971); Evolutionary Algorithm (n.d.), http://en.wikipedia.org/wiki/Evolutionary_algorithm

Good, I.: Speculations Concerning the First Ultraintelligent Machine. Advances in Computers 6 (1965)

Hanson, R.: Long-Term Growth As A Sequence of Exponential Modes (2000), http://hanson.gmu.edu/longgrow.pdf

Idel, M.: Golem: Jewish magical and mystical traditions on the artificial anthropoid. State University of New York Press, New York (1990)

Kahnemand, D., Slovicand, P., Tversky, A.: Judgement under Uncertainty: Heuristics and Biases. Cambridge University Press (1982)

Kurzweil, R.: The Singularity is Near. Penguin Group (2005)

Mallery, J.C.: Thinking about foreign policy: Finding an appropriate role for artificial intelligence computers. MIT Political Science Department, Cambridge (1988)

McCarthy, J., Minsky, M., Rochester, N., Shannon, C.: Dartmouth Conference. Dartmouth Summer Research Conference on Artificial Intelligence (1956)

Omohundro, S.: The basic AI drives. In: Wang, B.G.P. (ed.) Proceedings of the First AGI Conference. Frontiers in Artificial Intelligence and Applications, vol. 171. IOS Press (2008)

Ord, T., Hillerbrand, R., Sandberg, A.: Probing the improbable: Methodological challenges for risks with low probabilities and high stakes. Journal of Risk Research (13), 191–205 (2010); Paperclip Maximiser (n.d.), http://wiki.lesswrong.com/wiki/Paperclip_maximizer

Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approac. Prentice-Hall (1995)

Salomon, A.: When Software Goes Mental: Why Artificial Minds Mean Fast Endogenous Growth (2009), http://singinst.org/files/htdraft.pdf

Sandberg, A.: Friendly Superintelligence. Retrieved from Extropian 5 (2001), http://www.nada.kth.se/~asa/Extro5/Friendly%20Superintelligence.htm

Shulman, C.: Omohundro's "Basic AI Drives" and Catastrophic Risks, http://singinst.org/upload/ai-resource-drives.pdf

Shulman, C.: Whole Brain Emulation and the Evolution of. Retrieved from Singularity Institute for Artificial Intelligence (2010) http://singinst.org/upload/WBE-superorganisms.pdf

Simon, H.A.: The shape of automation for men and management. Harper & Row (1965)

Solomonoff, R.: A Preliminary Report on a General Theory of Inductive Inference. Cambridge (1960)

Steels, L., Brooks, R.: The Artificial Life Route to Artificial Intelligence: Building Embodied, Situated Agents (1995)

Sutton, R., Barto, A.: Reinforcement Learning: an Introduction. MIT Press, Cambridge (1998)

Turing, A.: Computing Machinery and Intelligence. Mind LIX 236, 433–460 (1950)
von Neumann, J., Morgenstern, O.: Theory of Games and Economic Behavior. Princeton University Press, Princeton (1944)
Walter, C.: Kryder's Law. Scientific American (2005)
Yudkowsky, E.: Creating Friendly AI (2001), http://singinst.org/CFAI/
Yudkowsky, E.: Friendly AI 0.9 (2001), http://singinst.org/CaTAI/friendly/contents.html
Yudkowsky, E. (n.d.). General Intelligence and Seed AI 2.3, http://singinst.org/ourresearch/publications/GISAI/
Yudkowsky, E.: The AI-Box Experiment. Retrieved from Singularity Institute (2002), http://yudkowsky.net/singularity/aibox