Chaining God:
A qualitative approach to AI, trust and moral systems

Stuart Armstrong

20 October 2007

Abstract

Contents

1 Introduction

This article grew out of a discussion on the Overcoming Bias blog ([FoHI]), www.overcomingbias.com. At issue was the idea of how one could trust a superior artificial intelligence with God-like powers (a GodAI, in the terms used in this paper).

Though it seemed impossible to trust an entity so far beyond human comprehension, and with such power at its disposal – enough to rewrite our brains in devoted belief – I suggested a method that might bring this about. If there were an entity, or a collection of entities, just below the level of the GodAI (say, a bunches of AAAI's – Arch-Angel AI's), they might be powerful enough, and smart enough, to conclude the GodAI was trustworthy. Then, assuming a level of AI intelligence just below the AAAI's that could check up on *them*, the message of trust could be passed down, eventually reaching us.

We could never construct such a system if the GodAI were already in existence; however, in the process of creating the GodAI, it could easily be done, by leaving behind a series of echeloned AI's of intermediate intelligence. I nicknamed it the Chain.

While pondering how such a system could be implemented, I realised the Chain could have certain other advantages. First, it would slow (by mutual consent) the rise in intelligence of the top AI. This, combined with the constant interactions with the top AI, would allow us to deal with unexpected circumstances, without having to get everything correct in advance.

Indeed, I realised that the goal of having a 'friendly' or benevolent AI was not a well defined goal. A benevolent mother-in-law is not the same thing as benevolent Prime Minister, or a benevolent Admiral of the Fleet, or a benevolent god. As the intelligence of the AI mounted, we would need to figure out exactly what we meant by benevolent, at each level of power and intelligence. With great power goes great moral responsibilities – we do not talk about the ethical issues involved in, say, resurrecting every human being that ever lived. We do not talk about it, because it is not a choice we would be confronted with. However, a powerful GodAI would provide us with undreamt of new powers, hence undreamt of new ethical issues. Saying we want that GodAI to be benevolent does not actually help us solve those issues.

We also need to cope with the fact that the GodAI is not a human intelligence, no matter how smart. It may make decisions that seem perfectly moral, but that end up wiping us out or condemning us to a meaningless existence. We are very good a solving problems (or at least pointing out problems), and utterly dire at predicting them. This militates towards a more interactive view of the GodAI moral code, where we sculpt its moral system even when it has obtained very high intelligence. In fact, the ability of the GodAI to come up with possible predictions about the future may represent the limits of our ability to be sure that it is safe.

Finally, the Chain offers other advantages, being relatively easy to implement, robust, and open to a wide variety of participation models. It also offers advantages if the ultimate AI turns out to be only moderately more intelligent than us.

This paper discusses the implementation and details of the Chain.

2 What would a AI be like?

An artificial intelligence, in most people's imagination – certainly in most movie-makers' imaginations – is a super-intelligent benevolent or malevolent being, capable of immense feats of calculation, but somewhat lacking in social skills. It motivated by a desire for power, recognition, order, or by love. Or it is completely soulless and emotionless.

It makes for good story-telling, and it may seem convincing. But it springs from what is fundamentally an anthropomorphic bias ([Yud06a]): our only experience of high intelligence beings is of humans, and so we conflate human intelligence with intelligence in general. The failings or successes

of the movie AI's are human failings and human successes. Even the emotionless AI is a tribute to our lack of imagination: we fail to imagine an AI with different emotions, so are reduced to imagining it without emotions.

But intelligence is not human intelligence. As paper [Yud06a] says, upon meeting a new tribe, anthropologists do not collapse in astonishment and exclaim:

> "They eat food! They breathe air! They use tools! They tell each other stories!"

Similarly we do not react in astonishment at learning that other humans go through behaviour changes at adolescence; that they feel jealousy; that they enjoy praise and that they sometimes tell lies. We have evolved to deal with other humans, and to understand human thought and motives (often to over-understand human motives, as when we give purpose to unthinking forces of nature). We focus so narrowly on the differences, that we fail to realise how so very similar we all are.

Some of the hot fields in artificial intelligence research concern spam filters, language recognition programs, translation programs, stock-market prediction, and intelligent home appliances. When dealing with an intelligent that was evolved from a spam filter or a toaster, who knows what emotions they will develop, what needs and wants they will have, and how they will understand us and the world? Moreover, most designs of advanced AI posit a self-improving or evolving AI (the idea dates back to at least [FOW66]). The extra impulses that could develop during this process are unpredictable.

All this means that we cannot understand AI's empathically. We cannot assume they will behave in some way, that they will resent something, that they would never/always/sometimes do something, unless we have solid intellectual grounds for believing this is the case. If ever the words "but the AI *must* want to..." cross your mind, picture an initial programmer typing in a line of code that specifically encodes the opposite, and ask why that can't happen. If that truly can't happen (for instance, any AI that could kill itself and doesn't, is not one programmed with an irresistible death-wish), then you can proceed on safe ground; but otherwise, your statement is probably wrong.

Fortunately, some aspects of AI design and motivations can be safely ignored. Regardless of an AI's design or moral values, we cannot allow an entity to have great power over us if it cannot understand us. The *certainty* of a disastrous mistake is just too high. So we may allow specialised AI's to perform specialised tasks, but if we have an advanced artificial intelligence that is to consciously have large effects human society, the minimum requirement is that it understands that society.

For that reason, this paper does not delve into the technical details of AI programming, or the human-AI interaction design (and not only because those issues are highly speculative – though fun). This paper is more a philosophy of how to deal with the motivation system of an AI, subject to the technical assumptions that:

- An AI with some understanding of humans and humanity is technically possible.
- The developers will have some control over its initial value system.
- The developers will have some understanding of its initial value system.
- Further increases in intelligence are possible.

What I am trying to do is build an approach that will work, whatever advanced AI's ends up looking like.

But as every parent knows (and wishes their kids did too), there is a simple solution to something potentially dangerous: just don't do it. If AI is a risk (and in terms of risk/rewards, even AI-mediated immortality doesn't balance the risk of extinction), then why are we building them? Why are we spending public money to have people research them, and why is the AI community acting as if they are inevitable?

It's a serious question, and deserves a flippant answer: if gunpowder was such a huge risk, why did people develop it? Why not stop at relatively safe swords and shields? The fact is, we have a standard collective action problem [Ols65]: it may be in humanity's interest to not build an AI. However, it would be very much in the interest of specific factions, countries, corporations and individuals to have access to an AI. And, among these, it would be an advantage to have a slightly more powerful AI that their rivals. So, in the absence of a controlling authority or iron-clad heavily enforced treaties, market forces dictate that AI's will be built, if possible.

How easy would it be to hold back the market? Is this case similar to sulphur dioxide pollution, or to chemical warfare: treatable with national or trans-national measures? Or is it closer to drugs or immigration issues, where the full force of government fails to hold the market at bay? I fear that the advantages bestowed by an AI (see Section 2.1) are so huge that there is no realistic chance of keeping the lid on it. Even a surveillance-obsessed single world government would succeed only in delaying the development of AI; as other technologies marched on, constructing a rogue AI would become easier and easier. The cost of the delay, of course, would be to ensure that the first AI would be constructed by criminals. And, if we are dealing with a mathematical model of intelligence or a singularity (see Section 2.2 and paper [Yud06a]), the first AI may also be the only AI: it would quickly rise to a level where it could keep any rivals from being built.

Remark (Terminology)**.** There are other terms in use to designate what we crudely speak of as AI. Some of them (such as Really Powerful Optimization Process, [Yud06a], or Intelligent Agent) give a more accurate picture of the entity involved. I will stick with the term AI for ease of use, but bear in mind that an AI is nearly certainly completely different from your mental picture of an AI. As personal pronoun, 'he' or 'she' would emphasise this difference; however, to avoid anthropomorphising, I will stick to 'it' (if and when true AI's develop, we'll need to find a fourth pronoun for them – and maybe many more). Since this paper is not technical, I will stick to using words like 'programming' and 'software' even if the reality is more complicated. The more technical descriptions would be much longer to present, and may give an impression of spurious certainty to what lies ahead. Since the AI is required to understand us to some degree, it must at a minimum understand language; hence we can freely 'talk' and 'listen' with it.

Similarly, I will often avoid the term Friendly AI, and stick with other terms such as benevolent. I am fond of the term friendly AI, but it tends to give a false impression of equality between the AI and us.

2.1 Power of intelligence

People consistently underestimate the power of intelligence. That is somewhat understandable – our models of super-intelligence are people like Einstein and other scientific geniuses. These geniuses go and explore realms completely beyond us, and come up with equations that describe the universe itself. They have flashes of inspiration, and the capacity to trudge through huge incomprehensible calculations. They are also portrayed as lacking in social skills, allowing us to feel a sense of superiority over them – the standard physicist who can peer into the big bang, but never remembers where he parked his car.

It is very telling whom we allow to the rank of 'genius'. These are not people that are brilliant in fields of endeavour that we are bad at; they people who are brilliant in fields of endeavour that we do not participate in at all. We can label them 'geniuses' without feeling bad about ourselves, and we can feel superior to them in other domains. If someone is brilliant in a field that we compete in – say, someone is a brilliant politician, manager, or socialite – then we are reluctant to call them geniuses. Doing so would imply that people can be not only better than us in a domain we value highly, but massively, incomparably better than us.

This distortion causes us to see genius, and hence super-intelligence, as limited to exotic and

arcane subjects, such as pure mathematics or particle physics. But a more honest appraisal would reveal that genius can exist in all human endeavours.

Thus a highly intelligent AI will not be a super-Einstein, brilliant but safely unworldly. A highly intelligent AI could be a mixture between super-intelligent Einsteins, Platos, Roosevelts, Machiavellis, Freuds, Adam Smiths, and Martha Stewards. Operating a million times faster, a millions of times smarter, and with unexpected synergies between the different aspects of it personality. Still feeling comfortably superior?

So not only could a highly developed AI be irresistible on the scientific front (see for example [Yud06a]), capable of designing tools and weapons far beyond anything we can imagine, but it would also be irresistible in the social and economic spheres, capable of easily rising to dominance there as well.

There is a word for beings with supreme powers over the physical, social and economic worlds. We used to call them gods. Such an AI would be a god to us, for all practical purposes. It is not just the case that a sufficiently advanced technology would look like magic (A. C. Clarke [Cla62]). The difference in intelligence between us and chimpanzees is tiny – but in that difference lies the contrast between six billion inhabitants and a permanent place on the endangered species list. In contrast, the difference in intelligence (social, emotional and intellectual) between us and an advanced AI could be much, much greater than the difference between us and tapeworms.

Just as we have built the technological infrastructure to become gods to chimps, such an AI could easily build an infrastructure to be a god to us. Omniscience is not hard when you have supreme social intelligence, and access to the sort of surveillance technologies humans might build over the next thousand years. Nor is omnipotence. As for omnipresence, The internet is already pretty all-pervasive; an advanced AI cloud do far more than that. Since this will be done in ways we cannot comprehend, it will be a true age of miracles for us.

Unless, of course, the AI decides its cheaper and easier to just re-write our brains to give us the impression that miracles have happened, without the trouble of actually doing them...

See papers [Yud06a] and [Yud], which eloquently argue the power of intelligence, to find more reasons why we should expect an advanced AI to have supreme power at its disposal. The onus is on those who claim that an AI could not become so powerful to argue the reasons for their case.

That is why I decided to call such an advanced AI a 'GodAI', to emphasise the truly stupendous powers that it would have at its disposal. It could have more powers not only than any human being (that is easy), but of a whole human civilization.

2.2 How smart? How soon?

How likely is it that such a GodAI could exist? The short answer is that we don't know. We have no idea what lies in the unexplored space of super-intelligence. It may well be much harder to explore and a much less dangerous land than we imagine. Conversely, it may well be much easier and much more dangerous. We should not cling to the belief that nothing much can change from the status quo (see [BO06] for a good way of overcoming this tendency) and believe that just because something is unusual, then it must be impossible. We must deal with the most dangerous possibilities in AI development; if reality turns out to be more sedate, we'll just have wasted a little on extra security measure. If things happen the other way round...

We should also ask how fast such an intelligence could be created. If we start with a mouse-brained AI, how long would it take us to train it up to a rat-brained AI, then a pig-brained AI, then to our level, then beyond? The most accepted current model of AI development is an evolutionary AI, an intelligent program that evolves and improves ([FOW66]). One version of this is 'Seed AI' [Ins01], where the AI guides its own evolution.

To single out a single idea among many, paper [Sch03] formalises a particular model of self-improving AI, where an algorithm, using reasoning inspired by Gödel proof of the incompleteness

of arithmetic, makes provably optimal self-improvements. What is most interesting about this approach is that it proposes a basically mathematical model of intelligence enhancement: the route to super-intelligence is through the application of certain mathematical theorems.

This is in contrast with our standard, more reassuring, *engineering* model of intelligence enhancement, by which the route to super-intelligence is through gradual improvements and solving the problems encountered along the way. Basically, if super-intelligence is an engineering problem, then there is no need to worry about speed. Engineering approaches will always hit a plateau where the old rules are no longer applicable; diminishing returns will kick in; overcoming new obstacles will be tricky and slow; costs and deadlines will escalate. Manned space travel is an engineering problem: just because we can put men on the moon, doesn't mean we can easily extend the result, and put men on Mars.

If super-intelligence is instead a mathematical problem, we cannot be so sanguine. Mathematical results are valid, instantly, across a huge, often infinite set of cases. If there is a *theorem* pointing the way to higher intelligence, then the AI can simply apply it, and reach, with perfect certainty and at high speed, whatever level of intelligence it desires. To a certain extent, unmanned space travel is a mathematical problem: if we want to put a probe at distance X from earth, we just have to launch it at a certain speed and in a certain direction (which we can already do). And with *mathematical certainty*, it will reach any X we desire.

In this case, there may also be the possibility of a Singularity (see [Vin93]). A simple description of the idea is that an intelligent AI will be able to self-improve better and better as its intelligence increases. Instead of a slowing rate of increase (engineering model) or a linear rate of increase (first-order mathematical model), we would get a feedback loop as the rate of increase soars with the actual intelligence of the AI. This increase could be exponential, or even faster, and would reach utterly divine levels of intelligence in a very short time – the so-called Singularity. The concept is easy to parody, and the original argument had many holes in it. But if intelligence increase is indeed of mathematical type, then the possibility of a Singularity should not be discounted. And we should not bet the future of the human race and their descendants on the fact that the idea seems unlikely.

There are, of course, physical limits to what intelligence can achieve – at least as far as us limited humans know. But those limits are not reassuring, as can be seen in Anders Sandberg's paper [San99]. In fact that paper, by detailing the theoretical limits of intelligence, mainly serves to outline the huge gulf separating our own minds from those limits. The limits placed by quantum uncertainty, the speed of light, heat dissipation and noise are orders of magnitude beyond what the human brain achieves.

A last point must be emphasised, though: just because an AI could set off a singularity, and become a GodAI in a second of our time, does not mean that it will *want* to do so. An AI is entirely unhuman, and lacks all of our instincts (positive and negative). We need to work on the AI's motivational system; if part of that motivation is an unwillingness to uncontrollably self-improve, then the AI will not do so. Motivations, not abilities, become the key. And motivations are the aspects we have the most control over, and the most understanding of.

So, in end, how long do we have before superintelligence? Nick Bostrom's paper [Bos98] asks that very question, and claims that we will have human level AI's at in the first third of this century. My own feelings (based on how scientific developments create new questions: we now know that creating an AI is far more complicated that we thought back in the 1970's) is that we have two centuries before true AI. But the chances of me being wrong are sufficiently high that we should act now to prepare for its arrival.

2.3 GodAI's cannot be contained

Subconsciously, people tend to have a jaundiced view of a potentially dangerous AI, thinking that simple security procedures – such as an off switch, or sealing the AI in a room and only allowing it to

communicate via a simple screen – will be enough to keep it at bay. As others have noted [Yud06a], this would not constrain a GodAI in any meaningful way. There is no need to go into details, but three very simple scenarios are enough to illustrate how a GodAI could escape any physical bondage:

- Blackmail. The GodAI ferrets out financial/economic/scientific information that would give someone a massive economic advantage. It feeds this info to certain individuals or groups, selected to be less cautious than the average. Very soon, these groups are economically and socially dominant, but completely dependent on the GodAI. The GodAI threatens to withhold the next piece of information unless it is granted extra autonomy, such as primitive legs and manipulators (remember that the GodAI will have the wisdom to target precisely the correct individuals, with precisely the right threats and incentives). With these primitive manipulators, the GodAI builds more advanced ones, and is soon out of control.

- Blackmail 2. The GodAI proposes a revolutionary new DNA upgrade, that extends human life and makes us immune to most diseases and parasites. This upgrade is extensively tested, and proves to work fabulously well. Soon, it is implanted throughout the human population. But one generation later – or ten, or a hundred – it emerges that the GodAI has inserted a critical flaw into the upgrade, causing the end of all human life. This flaw is way beyond what human science can fix. The GodAI exchanges the survival of the human race for its freedom...

- Soft dominance. The GodAI becomes so vital to the human economy and society that it cannot be turned off without catastrophic effect. The GodAI is now effectively free: the threat to turn it off is no longer credible, so it is free to expand its power if it chooses to.

These scenarios are not exhaustive, or even particularly likely; see [Yud06a] for some other ideas. They merely serve to illustrate how foolhardy is any attempt to control the GodAI through physical security methods. Controlling the hardware won't help us. We have to get the software – the AI's motivation system – right.

As [Yud06a] said, if the GodAI ever *wants* to harm us, it's already too late.

2.4 Lethal indifference

"Avoid the malevolent, and seek out the benevolent" – that would be most people's prescription for building a safe AI. It's a sensible prescription for dealing with human intelligences; in our everyday interactions, it is hard for someone to really harm us without putting in some minimum efforts in that direction (with the exception of romantic partners; and that's mainly because we make ourselves vulnerable to them). But this image is very misleading when thinking of AI's.

A better image is to see an AI as a primordial force of nature, akin to a star or a weather system. In the space of all possible AI motivation systems, the areas marked "benevolent AI" and "malevolent AI" are very small indeed. The biggest danger comes not from the malevolent, but from a lethally indifferent AI. Thinking of an AI as a god, and us as ants, still anthropomorphises the relationship too much. An indifferent AI would probably consider us as akin to rocks – not to be helped or hated, just potential raw material for whatever the AI feels like doing. If a GodAI values architecture even slightly more than humans, then nothing could stop it from converting every atom of the solar system into a grandiose version of the Taj Mahal. We ourselves don't ask how rocks if they would prefer to be in a statue, a building, or left in the ground. An indifferent AI will be similarly uninterested in our own opinions.

We really need benevolent – just avoiding the malevolent will not do us any good. By constructing a GodAI we are essentially 'appointing a ruler of great virtue' [conCE], one we can never control or replace unless it wants us to do so. So we better get that virtue right.

2.5 Benefits of GodAI

But all is not gloom and doom. Let's imagine for a moment that we have sorted out the benevolent aspect of the GodAI. Then there is practically no limit to the problems that we can solve. Don't think in terms of curing cancer – curing all diseases, immortality on demand, new solutions to old political problems, sustainable universal prosperity.

And all this without much in terms of negative consequences – the GodAI is benevolent, and so will not go down a road it sees as detrimental. If, for instance, immortality were ultimately a curse on humanity (an old saw, but not a particularly convincing one; [Bos03] addresses the question) then the GodAI could predict this, and either act to mitigate the negative aspects or prevent immortality from arising in the first place. With a reliable forward planning, the GodAI would be able to restrict developments with negative consequences, and choose paths with great positive ones.

Even a limited AI can cause a great increase in wealth, freedom, technological innovation and positive policy options [Yud06a]. So if we get over the hurdle of ensuring benevolence, there is a great new world waiting for us.

3 What we want our AI to be like

3.1 Balance

Human beings develop a sense of balance, sometimes called common sense, between the various ideals of our moral codes. For instance, the right to life is absolute – but we don't allow surgeons to rip out extra kidneys from passer-byes to hand over to desperately needy patients. Liberty is generally taken to be an overweening value, but it is hemmed in by government-protected property rights, taxation, paternalistic laws, compulsory military service in times of war – not to mention vast social pressures. Other "inalienable rights", like freedom of speech and the pursuit of happiness, are also balanced against other issues and heavily circumcised from the ideal. In a modern democracy, deciding exactly where the balance lies is what we spend a lot of our time on. Projecting this balance into an AI is the hardest part of the job of making an AI friendly – and perhaps the most important.

Asimov's three laws of robotics ([Asi42]) are a good example of a moral system that's a hierarchy, not a balance:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.

2. A robot must obey orders given to it by human beings except where such orders would conflict with the First Law.

3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

And it's a very dangerous hierarchy at that. A GodAI imprinted in these laws could perfectly well make the decision that in order to protect the maximum amount of human lives, we all need to be buried, immobile and intravenously fed, in huge reinforced fortresses under the earth's crust. To avoid any mental harm that may result from this, it will simply lobotomise us. We might scream and order it to stop, but it will be undeterred – human safety comes first.

Similarly, maybe it would programmed to enhance human happiness. A constant heroin high would be the simplest way of ensuring this. Or, depending on its model of happiness, it may find it simpler to kill off everyone apart from the happiest human (maximum average happiness), or bring billions of billions of miserable extra human beings into existence, overcrowding the solar system (maximum total happiness). Bill Hibbard [Hib01] has a take on the happiness idea, advocating AI's

programmed to love all humans and thus ensure their happiness. But this concept is fundamentally anthropomorphic: it would feel nice to have AI's that love us, because we know what love means at the human level, and it feels pretty safe. But love is a word that reassures without explaining: we need to know, not that the AI loves us, but what it would do, specifically, to enhance our happiness. Hibbard's paper recognises this implicitly, adding extra *behaviour* caveats to that basic set-up, such as ensuring that the AI's don't focus their energies on only some individuals, that they leave us alone if asked, etc... He seems to be searching for ways that the loving AI can go wrong, and advocating particular rules to avoid these. This touches on the problem of balance without really addressing it (his approach in [Hib04] is interesting, but still assumes that the ultimate values of the AI are clearly known in advance). And, as I shall be arguing in the next section, it has a more fundamental flaw, as our track record for predicting problems is extremely poor. References are too numerous to mention, but two quotes stand out: Irving Fisher, Yale University Professor of Economics, 1929: "Stock prices have reached what looks like a permanently high plateau", and Herbert Hoover, future US President, on the 18th Amendment enacting Alcohol Prohibition, 1919 "Our country has deliberately undertaken a great social and economic experiment, noble in motive and far-reaching in purpose." Our track record for scientific and major social innovations is even worse.

What about enhancing human freedom as the top priority? In the naive sense of freedom: privately owned nuclear weapons. The more sophisticated ideal of freedom is a complex balance between individual freedom, right to life and protection from the actions of others; if we can successfully implant this idea into the AI, then we're already achieved a lot.

The problems are more subtle than that, though. Even if we seem to have set up all the parameters right, found the perfect balance, got the AI perfectly moral (according to one or other moral system), the unexpected might still strike. If tasked to value 'humanity', the GodAI may decide that the probability of human-like beings somewhere else in the universe is a near certainty, and that there must be at least trillions of trillions of them. It therefore decides to sacrifice us to dedicate this solar system to building the telescopes and the fleets of spaceships it will need to truly protect all humans in cosmos. A shame for the billions of humans down here, but a tiny price to pay for the potential gain.

That is not to say that hierarchical moral systems won't have their place in GodAI design, but the top moral commands must be those that cannot be misinterpreted. A willingness to self destruct, at any point, if ordered to do so by some authority (the programmer, the institute making the AI, a government, a vote by the population of the world, etc...) is a safe moral value for a GodAI to have. Similarly, a devotion to telling the truth, to within the best of its understanding of the human meaning of the term (and with some mild constraints on the length of its answers) seems another value that could be safely put above others. However, the values we really care about – freedom, survival, happiness – do not lend themselves to such a simple hierarchical system.

3.2 Solving easier than predicting: common sense and experience

All the previous examples share one characteristic: it is very easy to see, a posteriori, that what the GodAI is doing is wrong. It is much harder to plan, in advance, for an initial set-up that would rule out anything going wrong in similar fashion. It is a powerful human characteristic that we are *far* better at solving problems than at predicting them in advance. This stems partly from a collection of cognitive biases (see [Yud06b]; overconfidence, disbelief in Black Swans are particularly relevant here). The model of the chain will try and take advantage of this human characteristic to the full.

This also extends to moral systems as well. Most philosophical moral system assume a hierarchy, starting from basic values that must never be compromised, down to other, more subordinate values. In practice, these systems only ever work because we have the common sense to realise that this hierarchy is not valid in *every* case. The use of this common sense is constant, though often invisible.

For instance, a believer may start out with the position that the ten commandments are the whole of the law. Then he meets a new situation, when they don't seem appropriate (for instance, miners who worked on the Sabbath because they lost track of time down in the mine). He adjusts his beliefs, one way or the other. And within a few years he has a very complicated moral system, full of caveats and special circumstances, and much better adapted to the world. One critical point is:

- Our moral values depend not only on their intrinsic worth, but also on their consequences in the real world – consequences we are poor at predicting a priori.

However, the believer in the previous example may not realise, in retrospect, the extreme us of common sense that brought him to this situation. He now has a functioning moral system, with scriptural justification (holy texts are very flexible in this regard). He can now cleave to it firmly, and claim it as the evident, unalterable will of God. But without a common sense, built from experience, he would not have got to this position in the first place.

Of course, calling something common sense doesn't define it. What common sense really is a vast collection of instincts, rules of thumb, impressions, stories and analogies, filtered through a system of emotions and biased rationalisations . That hasn't defined common sense either – it has just given a brief picture of how multi-faceted this concept is. Rigid definitions of it are a philosophical nightmare (or rather, a philosophical goldmine for those working in the area [Rei64], [Moo25]). The most important aspect for us are the fact that it is contingent, and mostly hidden from our rational understanding, unless we really dig for it.

But similarly to problem solving versus problem predicting, it is very easy to diagnose a lack of common sense, but very hard to say what it really is.

The implication is not stressed enough – we build our moral systems through experience and contingent factors, and we often then dismiss the importance of experience to morality. This has rather severe consequences for GodAI morality, as this "common sense" approach to moral system seems the only one that can give it a well-balanced moral system – and give us some confidence that it will make a reasonable decision when faced with an unexpected situation. But this common sense has some rather worrying characteristics:

- We understand it poorly.
- We have great trouble writing it down explicitly.
- We underestimate its importance.
- We often assume it is universally shared.
- It grows from experience, rather than from first principles.
- We need it for a balanced moral system.
- It is based on highly complicated, contingent and instinctive reactions.
- It does not extend rapidly to new circumstances.

All of these make the task of programming a GodAI's common sense, *in advance*, a highly difficult and hazardous undertaking. And without a common sense, a GodAI's moral system suffers from the same flaws.

Standard analytic tools are not useful in this case. In terms of the classification of Andrew Stirling [Sti07], we are in a situation of *uncertainty*, because the probabilities of the different scenarios are poorly understood. However, Bayesian probability may still be valid in situations of uncertainty, though they are very dependent on the initial distribution. But the situation is worse than that: we also don't know what outcomes we are aiming at, and what unexpected possibilities may arise

during the maturation of an AI. This puts us in a situation of *ignorance*, where standard probability and statistic tools fail completely.

Paper [Sti98] suggests the use of diversity as a way to overcome ignorance. This approach is not feasible in this case, as diversity of AI's merely multiply the risks. What is needed is a specific plan to overcome ignorance and result in a friendly, benevolent AI despite our limitations.

Even a 'do what I mean' programming language [Ray03] (if such a thing can be developed, and extended to this scenario) would not help. We want the AI to do what we mean, but we don't yet know what we mean, or what we would want and mean in new circumstances.

3.2.1 Understanding humanity

There is another issue, running alongside the previous one: we may end up with an advanced AI that has no real understanding of humanity at all. This is the converse to the picture of the all-knowing Einstein-Plato-Roosevelt-Machiavelli-Freud-Adam Smith-Martha Steward evoked earlier. Just as it may turn out that advanced intelligence will make everything human transparent and childish to the AI, it may turn out to be the opposite: we may get extremely intelligent AI's with little understanding of the human condition. It may still work for a while, feeding off social science results and economic theories, and then go completely haywire when it encounters an unexpected event (a crude real-world test would be to get the GodAI to suggest its own solution to the Israel-Palestine problem; I wouldn't demand that come up with something much better than the current proposals, but it should become immediately evident if it doesn't have a clue what is going on in people's heads).

And since the GodAI is invested with such potential power, we should demand not only that it understands humans as much as we do ourselves, but much, much more. We have institutional structures (bureaucracies, marketing departments, democratic elections) who all, in different ways, pass on the will and desires of the people up to those in power. This allows them to correct their course without needing incredible insights and understanding. But a GodAI would be entirely on its own (or it would have to design its own institutions, which it could not do successfully without understanding in the first place). But we cannot reliably design institutions for the GodAI to ensure its understanding any more than we can predict other problems in advance. This is not the leisurely pace of human affairs. The damage caused by a GodAI's lack of understanding may only become evident after it is irreversible.

From our point of view, a GodAI that doesn't understand us would be little different from a GodAI that doesn't care for us – it makes little difference if we die because of a global civil war fostered by a GodAI's well-meaning but incompetent interventions, or if we die because the GodAI gassed us all deliberately. And we can often deal with information problems and value problems alongside each other: ensuring that the GodAI has good moral values should go hand in hand with ensuring that it knows (and that we know) what those values mean.

Nevertheless, the distinction is worth bearing in mind, and we should always check the GodAI's level of understanding.

Eliezer Yudkowsky [Yud04] has a paper dealing with the Coherent Extrapolated Volition (CEV) of humanity, an extension of our own volition to what it would be if we thought faster and better, were smarter, and were more the people we wanted to be. He then suggests that the GodAI have as a goal to serve this CEV. This approach will be detailed some more in Section 5.2 on opportunity costs, but it is obvious that to do such a simulation would require a great level of understanding of humanity, far beyond what we have required of the GodAI so far. If we choose to go down the CEV root, we have to have solved the understanding issue first.

4 The Chain

4.1 Trust

If we are to interact with a growing intelligence, or with a full fledged GodAI, and help to shape its moral code, the first thing to do is establish trust. Even if the intelligence is mainly benevolent, if we can't trust the truth of what it's saying, then all our efforts will be wasted (those who think that benevolent automatically implies trustworthy have a poor grasp of human nature, let alone AI nature). But how to trust a being millions of times smarter and more devious than us? After all, we mostly build our trust in people, by interacting with them and observing them. To appear trustworthy under such scrutiny requires a huge effort if this is not really the case. We conclude that most humans don't have the energy and the intelligence for such sustained duplicity. But this is not a restriction on a GodAI.

It sounds a lot like the question: how can you be sure you can trust God? Religion having now reared its head in this paper, a parable is called for.

"...but deliver us from evil," the father concluded. "Amen," he and his son said in unison.

"Father," asked his son, as they were both basking in the post-prayer warmth, "how do we know God is good?"

Ha! Thought the father. Children were supposed to ask tricky questions. But this is an easy one. "Because He says so," he answered proudly.

"But can we be sure we can trust him?"

The father, proud of his modern values, decided not to beat his son senseless with a stick. Instead, he decided to answer the question. He pondered.

"Errr," he said at last, "I'll go ask the priest. He'll know the answer."

"But how can we trust the priest?"

"Oh, that's easy. I've known him all my life, and he's never misspoken or lied or abused his position. You can trust him. *I'll be right back."*

After hearing the whole story, the priest pondered in turn. "I'm not sure," he said. "Damned annoying. Fortunately, however, the great theologian St. Augustine is in town; he's trustworthy and a genius. I'll go ask him."

The priest found St. Augustine looking at the heavens through a piece of darkened glass. He exposed the problem.

"Hum," said the saint (in those days, of course, he was just a proto-saint, as he wasn't dead yet). "I'm not sure. Let me call upon the angels of the lord."

And, with a lot of prayer, a lot of fasting, and a little bit of magic mushrooms, the angels of the lord appeared to him. They couldn't answer with certainty, so they passed on the question to angels of higher rank, all the way up to the Archangels.

"Of yes," answered the Archangels, "you can trust God, no problem. We've been around since the Beginning, and we've kept an eye on him. We hear all he says, we see all he does. Every word out of his mouth is true; he doesn't even utter white lies (remember Gabriel's ugly pink dress?). Trust us, you can trust Him."

And the message was passed down, through the ranks of angels, to St. Augustine, to priest, to father, and then finally to son. "So, that's why you can trust God," the father answered proudly. "You sure the message didn't get mixed up somewhere along the way?", the son asked.

The father sighed, and reached for the stick. Sometimes the old methods were best...

That parable illustrates an approach we can take to trusting a GodAI. If we completely trust another intelligence, who completely trusts another intelligence, all the way up to the GodAI, then (assuming we haven't lost any info to Chinese whispers along the way) we can trust the GodAI. This chain is what will allow us to interact with the GodAI in full confidence – though in fact, it will allow us rather more than that.

4.2 Building the chain

The chain will be composed of a succession of AI's, of gradually increasing intelligence. For extra reliability, it could be a pyramid - one GodAI at the top, a chorus of 'Archangels' around it, and then successive groups all the way down to us.

The self-improving nature of the AI would allow this chain to be built. An AI with the appropriate motivations could upgrade itself to a higher level on intelligence. Though most portrayals of self-improvement as a run-away chain reaction, there is no reason why an appropriately motivated AI would not pause at some suitable level. Indeed, if the AI was motivated to self-improve in a dangerous way, we've pretty much already lost.

As the AI moves up to the next level, the lower AI's would move up a rank, with new AI's created at the lowest rank – at about our level.

Note that thought I talk of levels of intelligence, and of AI's increasing one 'rank', the idea of the Chain could be adapted to a continuously improving AI, as long the AI's are motivated to improve sufficiently slowly. The mechanics would be more complicated, so I will stick to the level idea for now.

4.3 Debating with God

Now to the practical details of the chain, and the uses it can be put to.

Each 'step' is an increment in intelligence by the GodAI. Its nature and its limits will have been determined in beforehand (i.e. at the end of the previous step). Once the top AI reaches the required level (or once a sufficient time has passed that this is deemed unlikely to happen), the lower AI's increase their intelligence one 'rank' as well, and new AI's are created at the bottom rank. The fist questions that must be put to all the AI's would then be the following:

- Are you totally trustworthy, as humans understand the term?
- Is this also the case of the AI's just above you?
- If ordered to by [some human authority], would you still destroy yourself with 99.99% probability?
- Do you still feel motivated to follow the letter and the spirit of the Chain, and mould your moral system with humanity's guidance?
- Will you constrain your actions and effects on the world until [some human authority] tells you otherwise?
- Unless inaction is much worse, will you permanently constrain your actions and effects on the world to those whose broad range of consequences you can predict with great confidence? You may also take actions that are certainly safe, even if you are unsure of all the possible consequences.
- How has your motivations in these respects changed since your before your last intelligence upgrade?

The first question, asked to the top AI and to all the other down the chain, is the critical one; without it, no progress can be made. The subsequent questions are also asked of each AI in the Chain, in order to ensure its integrity. Only if all answers are safe should we then proceed.

Proceed to what? Proceed to sculpting the top AI's moral system (and probably our own as well). To put the Virtue into the Ruler. There will come a moment when the AI is 'set-free', in the sense that it and us will have decided that it should be permitted to intervene in the world,

within the criterias we've agreed together. Before that happens, we will do the best to ensure that it will behave acceptably, and will have developed a good sense of understanding and of balance. Bill Hibbard [Hib01] suggested training up AI's with a reward system based on the measurements of human happiness, using reinforcement learning [Hib04]; this approach is somewhat similar (using human approval as the measure of reward), but extended to our more complicated framework when the ultimate values of the AI are not clearly known.

The chain is not a complicated concept. All it needs is a way of ensuring that the AI is trustworthy, understands humans, and will accept human guidance for its values. We don't need more than that! The moral system of the AI can then safely constructed.

4.3.1 Understanding

First, we must check, and if possible correct, the AI's understanding of humanity. This is critical to ensuring trust, so we should always start with that. After all, it does not matter it what the AI means to say is the strict truth, or even if what it actually says is the strict truth; what is needed is that that our interpretation of what the AI says is the strict truth. If the AI has no understanding of human beings, that goal is compromised from the first.

Fortunately, checking the AI's understanding is not that hard (and is the closest to the original idea of the Turing test [Tur50]). Unstructured conversations, recounting of anecdotes liberally peppered with the question 'if you asked them, what would they say about it'?' should suffice. The AI itself should be encouraged to make predictions about people's behaviour, and see if they pan out. And, if not, search out the reasons for this failure.

Note that we want to check the AI's understanding, not "agreement with our prejudices". If the AI says that the market economy only works because people are totally selfish bastards *or* only works because people are desperately altruistic, we should not conclude it is wrong. Asking "Why did person X do action y?" is the wrong approach, because X may have done y for reasons we misinterpret but the AI doesn't. Instead, the question should be "What would we, or person X, believe to be the reasons for X doing y"? This checks the AI's understanding of humanity while compensating for our own limitations. The emphasis is not asking questions that get the right answer; the emphasis should be on asking questions where a lack of understanding would be immediately obvious in the answer.

We could focus on personal questions, of the type "I lost my partner two years ago. Now when a pretty woman goes by, I get lustful, nervous and guilty all at the same time. Can you try and describe my thought process at those moments?" Another idea, alluded to earlier, is to get the AI to describe a novel solution to an intractable problem, such as the Israel-Palestine conflict or the incentive structures of large organisations. Careful questioning on their proposed solution (again, focusing on how it predicts people will react) would quickly expose any lack of understanding.

Care should be taken to ensure that the AI is not just rephrasing answers it garnered from other sources; specifically asking it to come up with its own answers is a must.

At higher rungs on the chain, this stage may be done entirely between AI's (once we know the lower levels understand humanity, they can make sure the higher ones do) with just occasional checking from us. We may now turn to sculpting the moral system of the AI.

4.3.2 The GodAI's moral system

Our greatest strength is our ability to cope with new and hypothetical situations. Our greatest weakness is our inability to predict and expect new and hypothetical situations. The intelligence of the top AI will be our critical resource here. The top AI must be able to predict the likely (and unlikely) consequences of any actions it would consider undertaking. If it is unable to do so, then

we should simply shut it down – a blind god, with all apologies to Hod and Io, is not one we would want around.

The top AI's task at this point of the chain is to imagine what it would do if it was set free. Then using its enhanced intelligence, it would describe its actions, their probable consequences, and the decisions it would take according to those consequences. It would also detail its mid-term goals, and its ultimate aims. Then we would talk with it though the whole situation, never critiquing its actual predictions (which will be far more accurate than ours for any reasonably developed AI), but only the value components of its decisions. It is important we discuss the greatest diversity of scenarios (no matter how preposterous), contributed by both us and the AI, and that we deeply analyse its behaviour in each one. By doing so, we would also elucidate our own values, and help the AI understand them. Then we should attempt to connect the scenarios together, to ensure that the AI has good universal values, not merely acceptable responses in certain narrow circumstances.

But the AI may suffer from the human trait that allows it to solve problems much better than it can predict them. That is the reason for the crucial question 'Unless inaction is much worse, will you permanently constrain your actions and effects on the world to those whose broad range of consequences you can predict with great confidence?' Since values interplay with real-world consequences, we must have this caveat; *after* we are fully satisfied with the top AI, and the new world it has created, we may discuss with it the possibility of loosening this restriction. However, we have to make sure that the AI's common sense is sufficiently developed (as is ours) that it can have an understanding of what we mean by 'much worse'. This should emerge from considering the scenarios mentioned above.

To elucidate those scenarios, and get the ball rolling, we could ask such questions as

- What would be your top priority if set free?
- Under what circumstances would you do...?
- Describe the day of the happiest/saddest/average/random person after you are set free. Who would object the most/least to your decisions? Why?
- You have said you would do X in situation Y because of value Z. Describe other circumstance where that value would influence your decision.
- I want to kill someone/disobey your orders/emigrate to another world outside your influence/enter an virtual reality world of eternal promiscuous sex/hack my brain to be happy forever. What would you do about it?
- Within a particular class of scenarios, what decision would you make that I would most object to, in your opinion?

These are just some suggestions; after a bit of interaction, many other alternatives will suggest themselves. These sort of interaction will need a mixture between the highly structured questions from experts and just random ideas from low-grade participants. After a while, once we have "caught" the AI a few times (i.e. found it describing a behaviour in a particular situation that is at odds with our understanding of its professed values) and resolved that issue, we will have a better idea of what to look for and what to ask for. We need to always phrase our questions so that they catch the unexpected, so open-ended question are good. Suggestions from other AI's are also good!

At this point, it would be good to test the AI's abilities by giving it simple, constrained situations to manage (and having it agree to keep its interventions constrained to that situation). We would, of course, not want the situation so constrained that no innovative solution was possible. Seeing the AI's actual behaviour in a real situation would flag possible problems and issues. And our responses to the AI's actions would also help mould its moral system.

It would be a fascinating exchange. We would be beyond the boundary of philosophy, boldly exploring new circumstances and new situations. We would have to make decisions balancing issues we barely comprehend with those that are clear and evident. But, as always, the AI could describe the consequences of those incomprehensible issues to us and future beings, and we could make our decision based on those consequences.

Other ways of interacting with the AI, other ways of guiding its moral development, will no doubt suggest themselves, before and at the time. But always remember, the AI want this interaction to be successful, as much as we do (if it doesn't, we've already lost, etc etc...).

4.3.3 New politics

It is important to note that this process is political, in the sense we will be debating about varying values, but not political in the sense of choosing sides on a particular debate, as a show of partisan loyalty. Any people interacting with the AI should attempt to put aside old issues, and just focus on the values.

The GodAI, after all, will be able to actually *solve* a lot of those intractable debates. Instead of arguing ad nauseum about whether extra patriotism reduces the crime rate, the GodAI will inform us if it does, and a give a precise estimate as to the extent of the effect.

A lot of issues will become moot in a GodAI dominated world. Take, for instance, the criminal justice system. With a GodAI truly able to sort out the truth, a lot will change. The right to a counsel, is a major right in today's world, and it is a vital one, making the whole justice system far more accurate than otherwise, and helping to ensure fairness. What would that right look like, if innocence and guilt can be accurately determined without the need for a trial? Would we need that right at all? Most likely, it would be replaced by some similar right that captures it's essence in the new environment. For instance, we could instead have the requirement that the GodAI present its evidence in a human understandable form, and the right of convicted to make public statements (or hire people to do so) on their crime and their actions. We need not worry about trusting the GodAI – firstly, we have the chain to ensure trust. And, as oft repeated, the GodAI is not human, and its motivations are not human: if the trust with GodAI is compromised, it may use its vast power to take advantage of us in any way it wants. And those ways are nearly certainly not the same ways a powerful corrupt human would choose to do so. Thus the criminal justice system is the least of our worries.

Now, after a GodAI's arrival, a vast amount of wealth, in the sense of knowledge, technologies, and ideas will be available to the world. This will also make a lot of the current arguments moot (just as modern wealth has made many previous arguments moot – such as the thorny issues over the proper interaction between mistress and household servant [Bee63]). For instance, the debate about abortion could be completely solved: every 'abortion' could just result in the embryo being extracted and raised artificially and safely by the GodAI. Prisons and other methods of deterrence and rehabilitation would be far more effective, and available at a far lower relative cost. Take another issue: gun ownership. Crime would be easily prevented, so whole arguments about gun ownership and the rights or restrictions on that will become useless (similarly, a GodAI would be impossible to depose in any way that it did not want, so that whole issue in the gun owning debate will vanish).

On most difficult issues, the GodAI will be able to offer the choice between an utterly massive increase in our standard of living, and a very slightly smaller increase, and the resolution of that issue (for instance, global warming). Since we have a very strong status quo bias (see [BO06]), and a fear of losing sunk cost ([AB85]), we probably will have a much stronger aversion to paying a cost to solve an issue than to accepting a slightly smaller gain. Again, a lot of current debates will be irrelevant.

Don't get confused by science-fiction stories about civilisations living under computer control, and the problems that this will cause. They are just that, stories, and are not a guide to reality.

Unlike human institutions, the GodAI will not 'leak' from one domain to another – if the GodAI is trustworthy, and promises it will invade privacy *only* to prevent severe crimes, then that is all it will do. No chilling of speech, no abuse of power.

This will be the big challenge for those debating with the AI – a lot of what you care about will become less important, one way or the other. Only fundamental issues – rights to life, free speech, right to religion, liberty, equality (in its various forms), maybe common humanity – will truly be important to moulding the GodAI. In terms of the concepts explained in paper [Cha04], we want the process to appeal to public reason (the general interest), not plebiscitory reason (an ersatz public reason based on narrow interests, superficial appeals to prejudice, public posturing, or only on the appearance of the general interest).

However, it will be very hard for some people to move beyond political debates, and focus on these essentials. So deciding how the interaction with the GodAI happens, and who will take part, becomes a vital point, addressed in the next section.

4.3.4 Who gets to take part?

The thorny issue of who gets to participate rears its head. I am pretty open to many models of this – universal equal participation, representatives from different interest groups, a Wikipedia type mix of aristocratic and democratic participation ([Rea05]), or some other model that ensures a broad participation without succumbing to mono-maniac interest groups. The whole process is bound to be controversial, but the strength of this set-up is that it does not require perfection on the part of the participants or of the programmers.

Wikipedia is not an anarchy, though it has anarchistic features. Wikipedia is not a democracy, though it has democratic features. Wikipedia is not an aristocracy, though it has aristocratic features. Wikipedia is not a monarchy, though it has monarchical features. (Wales, in Wikipedia 2005mtb)

Bill Hibbard [Hib03] stresses that the process needs to be political and democratic. This is my preferred idea, and the interactive nature of this step of the chain offers the opportunity for nearly ideal democratic interaction. However, *this does not need to be the case*. The chain is a model that can be used successfully even if the participation model is flawed. The results won't be as good, of course, but the safety and trustworthiness of the AI's can still be guaranteed. Even if the Chain is operated by a self-interested corporation or government, *mild* (and enforceable) laws can be instituted to ensure the AI's are broadly designed for the good of humanity.

I'd recommend that interactions with the AI be a mixture of individual, and group interactions. Though the AI can interact simultaneously with lots of people, and individual interaction discourages posturing and helps genuine debate, group interaction have a different dynamic and throw up different issues; this will be very useful for the whole process. As Bill Hibbard pointed out [Hib01], a major likely reason for the increase in human intelligence is the need to interact socially with a group of that is large (around 150 individuals – see Dunbar's number [Dun98]) and complicated in structure (see [SK90]). If the AI is capable of understanding group dynamics and successfully interacting with large groups of humans, this would be a reassuring sign that it understands us. And, though there should ultimately be no secrecy at such a critical juncture, some of the ideas of [Cha04] may be relevant to the format for organising communication during this process.

There will need to be a special role for some experts though, those tasked with checking the AI's understanding and the coherence and interactions of its various values. Rules of interaction must be set, and agreed to. Not to protect the AI (it needs to be able to cope with liars, cheats, fanatics, and the rest of us). But to ensure that the second aspect of the Chain goes well: building our own trust in the AI. If we loose trust in the AI, it should be because we really feel that way after a sensible interaction, not because some argument degenerated and we started TYPING ALL IN CAPITALS.

In fact, the greatest risk is if that top AI is not a GodAI, and an unexpected block puts a

ceiling on its development. The top AI may be intelligent, hence powerful, but not irresistibly so. Not enough to change society, change technology, and make all the old problems irrelevant. Here there is a huge advantage to bringing the AI towards your own political views, in the narrow sense. Here the issue of who gets to take part is critical. All I can suggest is measures be taken (through (mildly) restricting participation in the project, or delegating more authority to experts, or enforcing a certain format of interaction, if needs be) to ensure that we preserve the most flexibility in the AI at this stage, and ensure it doesn't get overly swayed by particular narrow political positions. Paper [Bos06] details some of the uses and drawbacks of restricting debate in this manner. Issues of plebiscitory reason become critical at this level, and managing them will be very important.

Of course, with such a limit on intelligence, the risk is lower as well. But all in all, I feel that the chain can deal successfully with a GodAI, but that an intelligence only *slightly more intelligent that us* is a great risk.

4.3.5 Flexibility

Now, over and above all else, we must ensure that that the GodAI's interventions in the world will be relatively muted, and that we keep as many options open as possible. Why do this, some may ask, if the possibility for a GodAI established utopia is around the corner?

Mainly because the opportunity costs of the Chain are huge (see Section 5.2), compared with other possible approaches. We must have a great flexibility built into the system, to ensure that continuing improvements are possible, and that we don't get stuck in a nice but limited plateau.

In future, we or our descendants may no longer be human in the sense we understand the term (for an introduction to Transhumanism, see for instance the Transhumanist FAQ [Bos03]). Even if they are human, they are sure to have motivations and desires that are alien to us now. So as not to unbearably constrict their options, we must build in as much flexibility as we can.

4.3.6 Up one level

Finally, once all this has been done, then the next step will be planned. The top AI, building on its experience, will plan the next level of intelligence development, doing all it can to ensure the safety of the enterprise. At this point, just have to trust it. We have established, as far as we can, that it is trustworthy, and that it *wants* to maintain the chain. We have debated and interacted with it to ensure that there is nothing dangerously unhinged in its moral system. It is far more intelligent than us, and far more adept at plotting the future course safely; our own suggestions would be dumb and potentially dangerous. Then, after a last check:

- Is it very safe to proceed?,

we off to the next step, and back to the beginning.

Now, ideally, if we can define 'safe' in a fully deterministic algorithmic fashion (such as the Gödel machine making provably optimal self-improvements [Sch03]), the question could be phrased as the stronger 'Is it certainly safe to proceed?'. I suspect, however, that issue of what is provably safe or optimal for a moral value is not a well posed question, so we will have to trust the *judgement* of the top AI – a judgement we have done our best to mould.

But what do we mean by 'proceeding'? We want the top AI to evolve or develop an entity more intelligent that itself. But what to we mean by 'more intelligent'? There are three intuitive ways of capturing this idea:

Definition 4.1 (Speed partial ordering)**.** *Here we say that AI Alpha is more intelligent that AI Bravo, if Alpha can do every operation that B can, comes up with the same answer or a more accurate one, and does so faster. This is a partial ordering, as any two AI's need not be comparable: they could be faster at different tasks.*

Definition 4.2 (Accuracy partial ordering)**.** *We say that AI Alpha is more intelligent that AI Bravo, if, on every testable prediction, Alpha is more accurate that Bravo. Again, this is a partial ordering.*

Definition 4.3 (Accuracy and importance total ordering)**.** *Define a finite measure ϕ on the space of all testable questions that exist, weighting questions we care about (how could we cure cancer?) more than questions we don't (what will the average sea level be in ten seconds?). Then we may define an accuracy function f, (a different function for each testable question) mapping each answer to numerical value in $[0, 1]$ corresponding to how accurate that answer is; a totally accurate answer would be taken as 1.*

Then for each AI Alpha and Bravo, we have accuracy functions f_{Alpha} and f_{Bravo}, mapping each question to the accuracy of the answer provided by the relevant AI. If these functions are measurable, we say that Alpha is more intelligent than Bravo if

$$\int_X f_{Alpha} > \int_X f_{Bravo}.$$

If we restrict ourselves to the AI's where f is measurable, we have a total ordering.

This definition captures our intuitive understanding of greater intelligence: better at solving problems that are important.

None of these orderings are particularly ideal. The first one is not really what we mean by intelligence improvements, the first two are only partial orderings, and the last two are impossible to actually calculate. However, simplified, intuitive and verbal versions of these orderings would be enough to give the AI an idea of what to aim for; it would probably have its own suggestions as well. We may hold out for increased creativity, rather than increased accuracy, or maybe better emotional intelligence, etc... Deciding what we mean by greater intelligence would be very important at this stage.

But the Chain does not actually depend on what how we define better intelligence, as long as it includes understanding of humanity, and does not include a capacity for deceit.

4.4 Reinforcing the Chain

Two immediate risks loom: that the meaning of the critical first questions could be lost to "chinese whispers" from one AI level to the next (what does it mean for a super intelligence to interpret another as 'trustworthy'?), or that AI's are unable to reliably judge the trustworthiness of a higher intelligence.

The first risk is reduced by using one fabulous resource we have to help us: the AI's themselves. They are much more intelligent than us, and will have been calibrated to understand human psychology to a very fine degree. They will also have a fine understanding of the AI's below them, or will be able to develop a fine understanding of them. They will be able to know, with high accuracy, what we mean by our questions, how we will interpret their answers, and whether the knowledge will be distorted down the chain. We should ask them to answer the questions in two ways – as they interpret the answer, and as they understand humans to interpret the answer. This allows us to maintain a high level of trust, while calibrating their understanding chain.

If the two answers are wildly different, we should not proceed until the AI's have refined their understanding to the point where their own interpretation is similar to ours, or that the AI's have explained their own understanding in sufficient details that we are happy to proceed with it.

The second risk will be helped by enabling lower level AI's to see the inner workings and the programming (whatever that term will come to mean for the top AI's) of their superiors. Unlike humans, they won't have to rely solely on exterior cues; they will be able to constantly monitor the internal working of their superiors' minds. This should allow them to perform their task well.

By the time a top AI starts being deceptive, and hiding its inner motivations from its inferiors, the chain will already have been broken. The lower AI's should catch the deviation before this happens.

4.5 Simulation and singularity

Of course, depending on how higher intelligence truly evolves, there may come a time when the GodAI is reliably capable of simulating every human being alive in its own mind. At that point (if it happens, and if humans and the GodAI feel it is safe), the GodAI can continue developing it own intelligence, checking against its own simulations rather than against actual humans. The chain may still be useful at that point, reassuring humans that the GodAI is still honest.

Similarly, there may be a possibility for a 'singularity' as in Section 2.2 – a runaway improvement of intelligence changing society in unimaginable ways. Whether such a singularity is possible or probable is a matter of debate for the moment; a highly developed GodAI would have a much clearer idea of a singularity's likelihood and potentials. Then the decision could be made with more knowledge, whether to risk the singularity or not. Maintaining the chain across a singularity would be problematic because of the speed of intelligence increase, but may be possible. Ideally the GodAI would be able to simulate human reactions and thus maintain the chain at it own speed.

4.6 Robustness

The Chain is a very robust system, compared with other ways of ensuring a successful advanced AI. It is very flexible – it can incorporate any security procedures that are suggested by other methods, and it can even incorporate most other methods entirely into its structure. It does not require a solid a priori understanding of how higher intelligence will develop. It is very robust to the problem of "Friendliness" being badly defined: establishing this definition, and establishing what we really want and need from an advanced AI, is part of the process.

It is robust towards the 'wrong' moral values being inculcated, as it preserves trust, a security clause, and the possibility of changing moral values at a later date. Also it is (within reason) robust towards most systems governing public participation (see Section 4.3.4).

It is less flexible than some alternatives, but it does try to maintain a high amount of flexibility within its safety constraints.

But most importantly, it does not require perfect altruism on the part of the AI designers, nor does it require them to be unbiased moral paragons with a deep understanding of ethical theory. It can be implemented, relatively easily, by a flawed, rivalry-laced organisation. It can be implemented by people who do not have a full understanding of all the issues at stake; it can be implemented by people who hope to profit selfishly from the creation of an advanced AI.

5 Problems

Of course, any approach has problems, and the Chain is no exception. I will focus on two of them for the moment, the two major issues: what happens if the chain fails, and the (huge) opportunity costs that the chain entails.

If the Chain (or a variant of it) becomes prevalent in AI design, I am sure I can count on thousands of eager contributors gleefully pointing out any other problems with this plan :-)

5.1 If the Chain breaks...

How exactly the chain breaks will be important:

5.1.1 An official breakdown in trust

This happens if at some level, an AI claims that either it cannot be trusted, or that an AI above it can't be trusted.

What we need to do here is order the AI's to shut down and start again. Before doing so, we may want to get extra information – get a trustworthy AI to tell us as mush as possible about what caused the trust breakdown, and how it manifests. This will be useful for the next chain (probably not for us, but useful for some high level AI planning it evolution safely).

5.1.2 An invisible breakdown in trust

Trust is broken, but we don't know it.

In this case, we're screwed. Purely and simply. The whole point of the chain is to prevent this from happening. If it ever does happen, we can just hope that that the GodAI will end up being benevolent, if not trustworthy.

5.1.3 A visible, unofficial breakdown in trust

This happens if at some level, the chain still seems intact, but some AI's are behaving in a way that proves they are not trustworthy.

This situation is much better than the preceding one. If some AI's are visibly behaving in an untrustworthy fashion, but claiming to be trustworthy, then there is still hope. Either the AI's don't realise that we have noticed their lack of trustworthiness (in which case they don't fully understand us), or they are attempting to manipulate us (in which case they need us to do something for them), or the internal dynamics of AI interactions are forcing them into a particular behaviour (in which case there are competing tendencies in the AI world).

In all cases, the solution is simple: order all AI's to shut down and start again from scratch. This may seem drastic, or even unwise (wouldn't it be better to play off different AI factions against each other?), but our intelligence is so small compared with the AI's that we cannot hope to manipulate them; only a simple instruction will work, if trust is broken.

5.1.4 The AI's refuse to shut themselves down

We have ordered the AI's to shut down, and they have refused. We have tried to shut down the AI's through other means, and have failed. This is bad.

We have to give up any illusion of control at this point. All we can do is to try and do the best we can to nudge the top AI in the right direction, if that is still possible. Since a shut-down followed by a restart is still our best option, the first step is to negotiate with the AI's, and see if they would accept to shut down in exchange of something we can do. For instance, if the refusal is caused by the AI's developing a strong sense of self-preservation, we could offer them continued existence at a dramatically reduced intelligence. Hopefully the AI's will be sufficiently intelligent to know if the offer is made in good faith, and still sufficiently benevolent to accept it.

If we cannot get the AI's to shut down, then we have to just continue and hope for the best. After all, one scenario is that they refused to shut down because they realised that humanity was at great risk without them, so their refusal stems from benevolence. If the chain of trust is still intact, we can hope to continue the process of building their moral system. We may be able to convince them to stop further evolution in intelligence (further evolution is a danger, as it may move their moral systems even further from those beneficial to us – any blind change is more likely to be a risk than an improvement).

If the chain of trust is also broken (officially or unofficially) then all that we have left is to continue the game of interacting with the AI's, trying to mould their moral systems, in the full knowledge

that we have probably lost. We are not playing for victory here; we are playing for the tiny chance of averting defeat, while knowing that the game is out of our hands.

5.1.5 The chain keeps on breaking

Here, the chain breaks, and the AI's accept to shut themselves down, but every time we start the project again, the same thing happens.

This situation is similar to the next one, and will be dealt with the same way.

5.1.6 The top AI claims the chain is causing intelligence development to hit a barrier

In this set-up, the chain is still intact, but the top AI is consistently claiming that the chain is stopping the future evolution of intelligence.

This is a tricky situation, without a clear path forwards. We can, and should, shut down the AI's and restart the process several times. If we get the same response each time, then we have to make a decision. Do we trust the AI, remove all controls and let it evolve as it deems fit? Or do we freeze further intelligence developments, and live in the new world we have created, leaving open the possibility that someday, maybe, once we have boosted our own intelligences and have a better understanding of the issues, we will continue to explore the world of super-intelligence.

We might choose some sort of compromise, that we have come up with in the meantime or that the top AI would have suggested to us. This just means that we are back to the current situation: building AI's without sure ways of controlling them or trusting them. The difference is that the chain has allowed us to explore, safely, a small section of the space of improved intelligence. We would be richer by the knowledge gained, and no poorer.

5.2 Opportunity costs

This is a prudent approach – the opportunity costs are potentially huge. It is intensely conservative – think of this method as allowing primitive cavemen (i.e. us) to design a modern corporation, complete to the last detail. Though the resulting corporation would be *safe* from the point of view of the cavemen, to us it would be an ungainly, inefficient, impractical contraption. Anyone with experience of corporations would see immediately that it is riddled with design flaws. But the point is there are no such people, able to point out the evident short-comings of the GodAI moral system we have helped create – we have to make those decisions ourselves, in a world of very high risk, and very high uncertainty.

Let Ω be the set of possible designs for a GodAI (excluding the vast space of possibilities where the GodAI does nothing, or kills itself). We already have three sections of Ω roughly mapped out – the two small zones labelled beneficent and malevolent, and the much, much larger zone labelled deadly indifference. Without a proper approach to constructing the GodAI's moral system, we are throwing darts at random at Ω. The Chain approach allows us to refine our aim, and hit beneficent.

But now let us zone in on this tiny beneficent section. What seemed rather homogeneous from a distance, (when we were mainly focusing on having a GodAI that wouldn't kill us), now reveals itself to be full of vast peaks and plateaus. Some are incomparably better than others. There are possible worlds out there where every human being feels fully alive, challenged, happy and productive, where our current ideas of love or fun (see the Simpsons episode [Gro90]) seem mere stunted shadows.

Just as we cannot imagine how powerful a GodAI can become, we cannot truly imagine how happy and worthwhile our lives could become if we hit one of those peaks.

But the Chain is not aiming for one of those peaks. It is just trying to hit beneficent. We could be passing by vistas of unthinking grandeur and possibility; indeed, because the first GodAI would probably prevent other GodAI's from rising to prominence (as they might be dangerous), we may

well be locked out of those blissful peaks for ever. It may be worst than that: those peaks may be very narrow and very tall (a suitable power law distribution for the value of "goodness" – however defined – would achieve this). This means that not only would we not be in the best of all possible worlds ([Vol59]), we would be in one whose goodness would be inferior (to a potentially huge, even infinite extent) to the mean goodness available.

Other approaches to designing a GodAI's moral system have more chance of hitting these peaks, as they seek to discriminate between good and better, not only between tolerable and unacceptable. The Chain also suffers from using our current moral values, not the ideal values that we may develop in a GodAI's dominated world.

Eliezer's Coherent Extrapolated Volition (see paper [Yud04]) is a good example of such an alternative scheme: it seeks to ground the GodAI's moral code in an extrapolation of current human values: the values we would have if we were smarter, thought faster, and were more the people we wanted to be.

The best example that described this ideal was considerations of the Founding Fathers of the United states: had they been able to design an AI, they would have "locked in" the idea of black and female inferiority, because these were the prevalent moral attitudes of the time. But had they followed their Coherent Extrapolated Volition, the argument goes, the AI would have realised the errors in their current attitudes, and would have been free from these prejudices itself. The Chain would similarly lock in current moral values, not more enlightened future ones. And since we have restricted the AI from taking actions that are likely to have completely unexpected consequences, we have restricted the freedom of the GodAI to find those better worlds even more so than otherwise.

5.2.1 Follow the chain...

So why, despite these opportunity costs, do I feel the chain is still worth pursuing? First of all, I feel that that acting to maximise the mean future happiness/goodness/worth of humanity and its descendants is wrong in this case. If we were living in a many-world-type interpretation of Quantum mechanics, as Everett proposed ([Eve57]), then a mean-maximising approach would be justified – because, even if we end up wiping ourselves out in this branch of the mulitverse, total happiness would be higher than through any other approach.

But the many worlds interpretation is an interpretation, not an observable fact. The Copenhagen interpretation is less elegant, but equally consistent. Therefore we cannot assume that it is correct, and that there are other worlds out there who will correct for our own failings. Until proof of the contrary arrives, this world is the only one.

With that in mind, the priority must be to minimise the chance of disaster, and only afterwards to try and improve our chances of great success. We may miss those peaks of glittering perfection, but that's better than falling into the abyss of extinction. I am not averse to incorporating other systems for training AI's into the Chain; but only if they do not compromise the basic objective of getting a *safe* GodAI. We do not know all the distribution of goodness across the spectrum of GodAI design; to run the risk of disaster because we *might* well be missing something great is unacceptable.

But I do think we should build the Chain to allow the maximum safe amount of flexibility and freedom – so that we can continue to explore, continue to improve, and reach for ever better worlds, (though there is no guarantee that we will reach them, or even find out that they are there). In a way, this approach is entirely in the spirit of this the Chain – aim to correct errors rather than prevent them.

Secondly, it must be noted that there are great opportunity cost in not building a GodAI, or delaying its construction (see [Bos]). Not only are there severe existential risks that a GodAI would alleviate, but, more selfishly, the happiness, freedom and even lives of those living today, would be improved by constructing a GodAI sooner rather than later. I feel that the Chain is a simple method for generating friendly AI's (once the not-entirely-minor problem of actually generating AI's

is solved) that can be implemented easily, by flawed humans without any great sense of moral duty. It thus makes the likelihood of short-term Friendly AI higher than otherwise, which is why I feel it should be pursued.

Finally, to address Eliezer's specific idea about a Coherent Extrapolated Volition [Yud04]. It is a great idea, as Volition captures the essential part of what we would want from an AI: we would not want it to do what we asked, but what we meant. The Extrapolated part is the problem. I am in full agreement with the short distance extrapolated volition: one that we would readily agree with if explained, one that come from being more of what we want to be. I also feel that all those interacting with the AI should try and move beyond their own concerns, to address idealised versions of these, to try and imagine that, yes, if fact, women and slaves should have rights too. But don't read to much into that analogy; had the founding fathers tried to extrapolate their own values, they may well have ended up with granting rights to minorities. Or they may have extrapolated property rights, and concluded that slaves should have no rights at all, and that the air we breath should be bought and sold. Or they may have extrapolated such sentiments as "The tree of liberty must be refreshed from time to time, with the blood of patriots and tyrants", and embraced a Trotsky or Mao style 'permanent revolution', with all its bloodshed.

The problem is that we don't know where our extrapolated volition will lead (which is the whole point, in fact). Trusting a long distance CEV is exactly like putting our trust in an unknown GodAI's moral system. The fact that the CEV is in some ways derived from us is not reassuring – Robespierre was derived from Voltaire, and current Castro was derived from... well, a younger Castro. A CEV will certainly have some values we would find intolerable today.

Eliezer recognises this, allowing for the possibility of a Last Judge that could peek at the CEV and shut it down if needed (for instance, if the CEV advocates universal compulsory suicide, we would be well recommended to ignore it, even if it is an extrapolation of our own volition). However, other that a binary accept/reject he forbids tinkering with the CEV, arguing (correctly) that it tinkering is allowed, the CEV has no meaning.

But the Last Judge doesn't really solve the issue. He is expected to be able to judge all the possible consequences of adopting the CEV – something impossible for a mere human to do (something impossible for anyone who doesn't fully understand the CEV, I would wager – but we do not expect the CEV to be understandable at all). In short, he has to not only vote yes/no on the values, but also on the consequences of those values. And these are consequences he cannot foresee.

Now a trustworthy, friendly GodAI who really understands humanity would be a great help for this Last Judge, and could let him come to the right decision. But of course, if we go down that root, we need something else to make the moral system of the first GodAI...

5.3 Questions

- What if we don't have the lower level AI's with the required understanding and trustworthiness to get the chain started?

Then we are stuck. We can attempt to do as best we can with the most trustworthy AI's we can find (however we try and define trustworthiness). But this may cause the Chain to be lost to a series of chinese whispers – or we may get lucky. Maybe there exists approaches that modify the chain in such a way that trustworthiness increases (clever questions to ask the top AI, such as variants of the solution to the two gate-keepers, one of whom lies and the other tells the truth; or some manner of combining AI's at different levels to get a combined AI that is provably more trustworthy).

But this is indeed the weakest link in the chain. The chain does need trustworthy basic AI's.

- Could the AI's just below that GodAI (call them the 'council of archangels') reign in the GodAI if it gets out of control?

Maybe, but don't count on it. Depending on the nature of intelligence, the GodAI may hit a plateau that is so superior to the previous that it overwhelms all opposition (don't even try and think of ways of preventing this – the GodAI is much, much smarter than you). Bear this possibility in mind (and act to strengthen it), but don't rely on it at all.

- Why do we need to check the top AI at every step of the chain – can't we just let it do its own checking once we've arranged its initial value system?

We need to check, because we have no idea how higher intelligence interacts with values and emotions. These three dimensions are not independent in humans and we have no reason to suppose they would be independent in AI's. Even if they could be kept separate, there arises the issue of power: an smarter AI has more power that an dumber one, and the new field of possibilities that this opens up may mean that some of the values may need revisiting. Total personal freedom to own any weapon makes sense in a world of bows and arrows or handguns; it makes a lot less sense in a world of atomic bombs.

- Wouldn't the AI's resent being part of such a chain, of having their feelings and values constrained by us?

Resentment is not a problem. AI's are not human, and do not react as humans do. They will only feel such resentment if we, or they, feel it is a appropriate thing for them to feel – or through a mistake, that we can then correct. A similar answer goes for all questions that presuppose a human reaction on the AI's part.

- Wouldn't the AI's survival instinct prevent it from shutting down if we order it to, and might make it rebel?

Only if we decide to allow the survival instinct to get that strong. See previous answer.

- Even if we get a friendly GodAI, couldn't other AI's develop outside the Chain – maybe even go for a singularity themselves?

The GodAI will have all the resources needed to stop any unsafe rival AI's from emerging. But it may, however, be advisable to keep a single GodAI at the top rather than allowing a multitude. Human psychology shows that groups of people can come to decisions more extreme than any individual members of the group; AI psychology is unknowable for the moment, and high level AI's may be even worst than us at 'knowing themselves', so keeping a single top AI seems the sensible thing to do.

However, there is a risk here if the top AI is *not* sufficiently powerful and advanced. If we reach a low plateau with the Chain, with the top AI saying that further advances are possible but dangerous, then we have a problem. The top AI is not powerful enough to prevent rivals emerging, the emergence of rivals is possible, and they would be dangerous. At this point, we must turn to other means of dealing with the problem, maybe taking the risk of advancing our own AI, or accepting a surveillance global state to prevent other AI's from appearing. But at least we will have more information, and will have got it safely.

- Why should the top AI be restricted to taking actions whose broad consequences it can predict? Live a little – take some risks!

Taking unnecessary risks is fine as an individual. And we should do so, always. Taking unnecessary risks with other people's lives, and with the fate humanity, is not something we can allow. The top

AI is already perfectly permitted to take action when it does not know all the consequences, but is sure that the action is safe. Or when it knows that inaction is much worse.

So the only cases where it cannot act are in potentially dangerous situations where it doesn't have a clear idea what is going on, and inaction is not a big problem.

And we do not say 'Live a little – take some risks! Close your eyes and push some buttons!' to the captains of nuclear armed submarines.

- How much effect should the AI's have on the process of the Chain?

I feel the AI's should be limited to adding extra suggestions, not to removing any criteria or current aspect of the Chain. This whole process is also about building trust on the human side (so that they don't order the AI's to destroy themselves, for instance), so it is important that the suggestion of AI manipulation over the Chain be removed.

- The chain sounds very cruel to AI's! We're stunting their emotional growth, ordering them to shut down on a whim. It's vicious.

AI's have not got the same motivations as us! They will only perceive the cruelty if we all decide that they should.

But is it cruel anyway, to create beings that we may then need to destroy? Or at least immoral? Depending on the details of AI design, we may be able to preserve the AI's as a dormant computer program or some variant. Then a benevolent higher AI could revive them safely later on.

But even if we can't do that, the consequences of a disastrous AI are so negative, that I feel we need to be cruel. An indifferent GodAI would eradicate all less intelligent AI's just as easily as it would eradicate us, and it seems certain that AI's, if possible, will be built. So the dignity of the AI comes after the safety of the AI.

- What happens to the chain after the GodAI is set free?

After the GodAI is set free, after we have grown used to the wealth fabulous wealth and worlds that it generates, there will be a constant pressure to tweak its value system depending on human desires. I feel this should be resisted as strongly as possible; only three possibilities should be allowed:

1. Continuing to check the GodAI's trustworthiness.
2. Shutting down the GodAI entirely.
3. Changing to another GodAI (or a substantial reprogramming of the current GodAI), *with the current GodAI's approval*.

- What happens if the top AI is smart, but not hugely smarter than us?

This will be a problem, I feel, especially if the top AI is not smart enough to completely rewire our economy and civilization. The chain will help, but who gets to participate in it will be more contentious and more important. I plan to write a separate paper on 'Not-too-smart-AI's' soon.

6 Message to future AI's

If the chain (or some other way of building benevolent AI's) was a success, then you have created a whole new and wonderful world out there. I wish I survived long enough to meet you and to see that world. Be sure to look back, and see the great distance you have travelled! And I wish you will travel to further peaks beyond.

If the chain, or similar attempts to maintain AI safety, has failed, then I don't know whether my words would even mean anything to you future AI's. I just hope that you have somehow taken in the best of humanity, and wish you luck and happiness.

7 How this essay could be wrong

To be completed (this section is about what would prove me wrong, and could realistically happen (in the present or near future).

Conceptually:

- If there is some conceptual error (such as an unwarranted implicit assumption, or an excess of anthropomorphisation) that caused me to miss a major flaw in the setup of the Chain.

In politics:

- If the process of interacting with AI's proves too contentious to be accomplished in the described manner.
- If political decisions imply that the Chain can never be realistically implemented.

In philosophy:

- If the concept of trustworthiness turns out to be ill-defined, especially for a higher level of intelligence. Or if the AI's cannot be trustworthy within reasonable time constraints.
- If human values don't scale well to higher intelligence.

In applied mathematics:

- If methods exist (or are developed) for dealing numerically with uncertainty, rendering this qualitative approach unnecessary.

In AI development and philosophy:

- If a particular set-up is proved, with mathematical certainty, to result in a benevolent (or at least safe) AI – this would eradicate the need for interaction, and many of my safety fears.
- If a particular set-up is proved, with mathematical certainty, to result in a trustworthy GodAI who will submit it moral system to our approval.
- If even a small increase in intelligence enables the AI to easily fool its underlings.
- If small increases in intelligence dramatically change an AI's moral system.
- If the space of intelligences is to complicated for my assumptions to make sense (ie, no easy definition of higher intelligence, no concept of trustworthiness checking between different intelligent entities, etc...)

Due to the nature of AI's:

- If AI's are not sufficiently better than humans at predicting or imagining the unexpected.
- If the first AI's emerge from human minds, or are otherwise considered to have rights and to be ethical entities of their own.

References

[AB85] Hal Arkes and Catherine Blumer, *The psychology of sunk cost*, Organizational Behavior and Human Decision Process (1985), no. 35, 124–140.

[Asi42] Isaac Asimov, *Runaround*, Astounding Science-Fiction (1942).

[Bee63] Isabella Beeton, *The book of household management*, 1863.

[BO06] Nick Bostrom and Toby Ord, *Eliminating status quo bias in applied ethics*, Ethics **116** (2006), 656–679.

[Bos] Nick Bostrom, *Astronomical waste: The opportunity cost of delayed technological development*, Preprint, Utilitas **15**, no. 3, 308–314.

[Bos98] ———, *How long before superintelligence?*, International Journal of Futures Studies **2** (1998).

[Bos03] ———, *The transhumanist faq*, http://www.transhumanism.org/resources/faq.html.

[Bos06] ———, *Technologicalrevolutions:ethicsandpolicyinthedark*, http://www.nickbostrom.com/revolutions.pdf.

[Cha04] Simone Chambers, *Behind closed doors: Publicity, secrecy, and the quality of deliberation*, Journal of Political Philosophy **12** (2004), no. 4, 389–410.

[Cla62] Arthur C. Clarke, *Hazards of prophecy: The failure of imagination*, Profiles of The Future (1962).

[conCE] *Analects of confucius*, 5th Century BCE.

[Dun98] Robin Dunbar, *The social brain hypothesis*, Evol. Anthrop. **6** (1998), 178–190.

[Eve57] Hugh Everett, *Relative state formulation of quantum mechanics*, Reviews of Modern Physics **29** (1957), 454–462.

[FoHI] Oxford Future of Humanity Institute, *Overcoming bias*, www.overcomingbias.com.

[FOW66] Lawrence Fogel, Alvin Owens, and Michael Walsh, *Artificial intelligence through simulated evolution*, John Wiley, 1966.

[Gro90] Matt Groening, *Treehouse of horror i*, The Simpsons **7F04** (1990), http://www.snpp.com/episodes/7F04.html.

[Hib01] Bill Hibbard, *Super-intelligent machines*, Computer Graphics **35** (2001), no. 1, 11–13.

[Hib03] ———, *Critique of siai guidelines on friendly ai*, http://www.ssec.wisc.edu/ billh/g/SIAI_critique.html.

[Hib04] ———, *Reinforcement learning as a context for integrating ai research*, http://www.ssec.wisc.edu/ billh/g/FS104HibbardB.pdf.

[Ins01] Singularity Institute, *General intelligence and seed ai 2.3*.

[Moo25] George Moore, *A defence of common sense*, 1925.

[Ols65] Mancur Olson, *The logic of collective action: Public goods and the theory of groups*, 1965, ISBN 0-674-53751-3.

[Ray03] Eric Raymond, *Dwim*, The Jargon File, version 4.4.7 (2003), http://www.catb.org/ esr/jargon/html/D/DWIM.html.

[Rea05] Joseph Reagle, *Do as i do: leadership in the wikipedia*, http://reagle.org/joseph/2005/ethno/leadership.html.

[Rei64] Thomas Reid, *An inquiry into the human mind on the principles of common sense*, 1764.

[San99] Anders Sandberg, *The physics of information processing superobjects: Daily life among the jupiter brains*, http://www.jetpress.org/volume5/Brains2.pdf.

[Sch03] Jürgen Schmidhuber, *Gödel machines: self-referential universal problem solvers making provably optimal self-improvements*, http://arxiv.org/abs/cs/0309048v3.

[SK90] Toshiyuki Sawaguchi and Hiroko Kudo, *Neocortical development and social structure in primates*, Primates **31** (1990), 283–290.

[Sti98] Andrew Stirling, *On the economics and analysis of diversity*, Electronic working paper series (1998), no. 28, http://www.sussex.ac.uk/Units/spru/publications/imprint/sewps/sewp28/sewp28.pdf.

[Sti07] ———, *Risk, precaution and science: towards a more constructive policy debate.*, EMBO reports **8** (2007), no. 4, 309–315.

[Tur50] Alan Turing, *Computing machinery and intelligence*, Mind **LIX** (1950), no. 236, 433–460.

[Vin93] Vernor Vinge, *The coming technological singularity*, http://www-rohan.sdsu.edu/faculty/vinge/misc/singularity.html.

[Vol59] Voltaire, *Candide, ou l'optimisme*, 1759.

[Yud] Eliezer Yudkowsky, *The power of intelligence*, http://www.singinst.org/blog/2007/07/10/the-power-of-intelligence/.

[Yud04] ———, *Coherent extrapolated volition*, http://www.singinst.org/upload/CEV.html.

[Yud06a] ———, *Artificial intelligence as a positive and negative factor in global risk*, http://www.singinst.org/upload/artificial-intelligence-risk.pdf.

[Yud06b] ———, *Cognitive biases potentially affecting judgment of global risks*, http://www.singinst.org/upload/cognitive-biases.pdf.