Ben Goertzel

# *Should Humanity Build a Global AI Nanny to Delay the Singularity Until It's Better Understood?*

**Abstract:** *Chalmers suggests that, if a Singularity fails to occur in the next few centuries, the most likely reason will be 'motivational defeaters' — i.e. at some point humanity or human-level AI may abandon the effort to create dramatically superhuman artificial general intelligence. Here I explore one (I argue) plausible way in which that might happen: the deliberate human creation of an 'AI Nanny' with mildly superhuman intelligence and surveillance powers, designed either to forestall Singularity eternally, or to delay the Singularity until humanity more fully understands how to execute a Singularity in a positive way. It is suggested that as technology progresses, humanity may find the creation of an AI Nanny desirable as a means of protecting against the destructive potential of various advanced technologies such as AI, nanotechnology and synthetic biology.*

## Introduction

I find myself in almost total agreement with Chalmers' (2010) careful analytical treatment of the 'Singularity hypothesis'. However, as an intentionally high-level treatment, it leaves many critical points unelaborated; my goal here is to enlarge on one of these, namely the potential for what Chalmers calls 'motivational defeaters' for the

Correspondence:
*Email: ben@goertzel.org*

*Journal of Consciousness Studies*, **19**, No. 1–2, 2012, pp. 96–111

transition from what he calls AI+ (extensible, human-level AI) to what he calls AI++ (Singularity-constitutive, dramatically superhumanly intelligent and powerful AI).

Specifically, I will elaborate on one possible motivational defeater: an 'AI Nanny', defined as an advanced AI+ system explicitly designed to thoroughly surveil the Earth and keep humanity safe from various dangers, including unpredictable advanced technologies such as AI++. AI Nannies could potentially be designed to forestall Singularity forever, or alternately to slow down the path to Singularity in order to enable it to proceed in a more deliberate way with a higher chance of a positive outcome. I will discuss the plausibility of creating AI Nannies, and also the ethical desirability of doing so – which I conclude is ambiguous, but potentially positive.

## Chalmers on Motivational Defeaters

Chalmers' argument for the plausibility of a Singularity occurring in humanity's mid-term future is, compactly, that

> an intelligence explosion results from a self-amplifying cognitive capacity, correlations between that capacity and other important cognitive capacities, and manifestation of those capacities (conclusion). More pithily: self-amplification plus correlation plus manifestation = singularity.

He then considers classes of possible 'defeaters' that might cause his argument for the Singularity's occurrence to fail, observing that

> We can divide the defeaters into motivational defeaters in which an absence of motivation or a contrary motivation prevents capacities from being manifested, and situational defeaters, in which other unfavorable circumstances prevent capacities from being manifested.

And he notes that the motivational defeaters seem perhaps the most likely to arise in practice:

> Speaking for myself, I think that while structural and correlational obstacles (especially the proportionality thesis) raise nontrivial issues, there is at least a prima facie case that absent defeaters, a number of interesting cognitive capacities will explode. I think the most likely defeaters are motivational. But I think that it is far from obvious that there will be defeaters. So I think that the singularity hypothesis is one that we should take very seriously.
>
> …
>
> … [I]t is certainly possible that AI+ systems will be disinclined to create their successors, perhaps because we design them to be so

> disinclined, or perhaps because they will be intelligent enough to realize that creating successors is not in their interests. Furthermore, it may be that AI+ systems will have the capacity to prevent such progress from happening.
>
> …
>
> A singularity proponent might respond that all that is needed to overcome motivational de- featers is the creation of a single AI+ that greatly values the creation of greater AI+ in turn, and a singularity will then be inevitable. If such a system is the first AI+ to be created, this conclusion may well be correct. But as long as this AI+ is not created first, then it may be subject to controls from other AI+, and the path to AI++ may be blocked. The issues here turn on difficult questions about the motivations and capacities of future systems, and answers to these questions are difficult to predict.

Here I will follow up on these points – specifically, the potential that humans might design AI+ systems to be disinclined to give rise to AI++ systems (in spite of possessing the capability to do so), and to also prevent other AI systems from achieving the capability to make the transition to AI++. I will not attempt to argue for this scenario's likelihood, but merely for its plausibility. I will also address the ethical question of whether an AI+ 'AI Nanny' of this sort is a desirable . A key point I will emphasize is that, even if you believe a Singularity is a laudable ultimate goal for humanity, it may still be rational for you to favor the creation of an AI Nanny with a predetermined finite lifespan, with the goal of mediating a slower and more reliably positive path to Singularity. If a sufficient number of sufficiently powerful humans conclude that the creation of an AI Nanny is ethically desirable, this obviously may increase the probability of an AI Nanny scenario coming about.

## The Ethical Motivation for an AI Nanny

In his paper, Chalmers carefully considers the question of whether a Singularity may occur, but avoids the issue of whether, if it does, this will be a good thing or not. While this is an understandable omission, resulting in a more compact and simple and less controversial treatment, it's also important to consider that human perceptions of the ethical character of the Singularity are plausibly likely to play a significant role in how (and if) the Singularity unfolds. Here I will come at the AI Nanny idea from the direction of ethics — as a potential (partial) solution to the moral dilemmas the Singularity poses.

The dramatic potential of a Singularity for both ‘good’ and ‘bad’, according to folk morality standards, is fairly obvious. The ongoing advancement of science and technology has brought us many wonderful things, and will almost surely be bringing us more and more, even before the occurrence of a full-on Singularity. Beyond the ‘mere’ abolition of scarcity, disease and death, there is the possibility of fundamental enhancement of the human mind and condition, and the creation of new forms of life and intelligence. Our minds and their creations may spread throughout the universe, and may come into contact with new forms of matter and intelligence that we can now barely imagine.

And on the dark side, Nick Bostrom (2002) has enumerated some of the ways that technology may pose ‘existential risks’ — risks to the future of the human race — as the next decades and centuries unfold. And there is also rich potential for other, less extreme sorts of damage. Technologies like AI, synthetic biology and nanotechnology could run amok in dangerous and unpredictable ways, or could be utilized by unethical human actors for predictably selfish and harmful human ends.

Given this tenuous balance of benefits and dangers, and the plausible likelihood of a Singularity occurring, it’s understandable that many are disturbed by our almost total lack of understanding of the odds of the various possible outcomes ensuing from the Singularity. And this train of thought leads to the perspective that intentionally creating a ‘motivational defeater’ to forestall or at least delay the Singularity might be ethically advantageous. Wallach and Allan (2008) review these issues and conclude that a Singularity is sufficiently far off that we should currently focus our ethical attention on nearer-term problems; but, my own view is that our uncertainty about future research progress is sufficient that it behooves us to take the possibility of more rapid progress toward Singularity very seriously.

The possibility I wish to explore here is the creation of a powerful yet limited AGI (Artificial General Intelligence) system (an AI+ in Chalmers’ terms), with the explicit goal of keeping things on the planet under control while we figure out the hard problem of how to create a probably positive Singularity. That is: to create an ‘AI Nanny.’

The envisioned AI Nanny would forestall a full-on Singularity for a while, restraining it into what Max More (2009) has called a Surge, and giving us time to figure out what kind of Singularity we really want to build and how. It’s not entirely clear that creating such an AI Nanny is plausible, but I’ve personally come to the conclusion it probably is. It's also not entirely clear that, even with the help and supervision of a well-built AI Nanny, humanity would ever understand the Singularity well enough to feel comfortable moving forward with it — which highlights the question of how long an AI Nanny would be empowered, if it were created.

### Perspectives on the Ethical Dilemma of the Singularity

Given the ethical complexities mentioned above, to which I've suggested an AI Nanny might be a possible solution path, what does the contemporary pantheon of futurist gurus think we should do in the next decades, as the path to Singularity unfolds?

Kurzweil (2005) has proposed 'fine-grained relinquishment' as a strategy for balancing the risks and rewards of technological advancement. But it's not at all clear this will be viable, without some form of AI Nanny to guide and enforce the relinquishment. Government regulatory agencies are notoriously slow-paced and unsophisticated, and so far their decision-making speed and intelligence aren't keeping up with the exponential acceleration of technology.

Further, it seems a clear trend that as technology advances, it is possible for people to create more and more destruction using less and less money, education and intelligence. There seems no reason to assume this trend will reverse, halt or slow. This suggests that, as technology advances, selective relinquishment will prove more and more difficult to enforce. Kurweil acknowledges this issue, stating that 'The most challenging issue to resolve is the granularity of relinquishment that is both feasible and desirable' (*ibid.*, p. 299), but he believes this issue is resolvable. I'm skeptical that it is resolvable without resorting to some form of AI Nanny.

Eliezer Yudkowsky (2002; 2009) has suggested that the safest path for humanity will be to first develop 'Friendly AI' systems with dramatically superhuman intelligence. He has put forth some radical proposals, such as the design of self-modifying AI systems with human-friendly goal systems designed to preserve friendliness under repeated self-modification; and the creation of a specialized AI system with the goal of determining an appropriate integrated value system for humanity, summarizing in a special way the values and aspirations of all human beings (Yudkowsky, 2004). However, these proposals are extremely speculative at present, even compared to feats like creating an AI Nanny or a technological Singularity. The practical realization of his ideas seems likely to require astounding breakthroughs in mathematics and science — whereas it seems plausible that human-level AI, molecular assemblers and the synthesis of novel organisms can be achieved via a series of moderate-level breakthroughs alternating with 'normal science and engineering.'

Bill McKibben (2004), Bill Joy (2000) and other modern-day techno-pessimists argue for a much less selective relinquishment than Kurzweil. They argue, in essence, that technology has gone far enough — and that if it goes much further, we 'legacy humans' are bound to be obsoleted or destroyed. They fall short, however, in the area of suggestions for practical implementation. The power structure of the current human world comprises a complex collection of interlocking powerful actors (states and multinational corporations, for example), and it seems probable that if some of these chose to severely curtail technology development, many others would *not* follow suit. For instance, if the US stopped developing AI, synthetic biology and nanotech next year, China and Russia would most likely interpret this as a fantastic economic and political opportunity, rather than as an example to be imitated.

Hugo de Garis agrees with the techno-pessimists that AI and other advanced technology is likely to obsolete humanity, but views this as essentially inevitable, and encourages us to adopt a philosophical position according to which this is desirable. In his book *The Artilect War* (2004) he contrasts the 'Terran' view, which views humanity's continued existence as all-important, with the 'Cosmist' view in which, if our AI successors are more intelligent, more creative, and perhaps even more conscious and more ethical and loving then we are — then why should we regret their ascension, and our disappearance? In more recent writings (2011), he also considers a 'Cyborgist' view in which gradual fusion of humans with their technology (e.g. via mind uploading and brain computer interfacing) renders the Terran vs. Cosmist dichotomy irrelevant. In this trichotomy Kurzweil falls most closely into the Cyborgist camp. But de Garis views Cyborgism as largely delusory, pointing out that the potential computational capability of a grain of sand (according to the known laws of physics) exceeds the current computational power of the human race by many orders of magnitude, so that as AI software and hardware advancement accelerate, the human portion of a human-machine hybrid mind would rapidly become irrelevant.

Considering these views all together, the dilemma posed by the rapid advancement of technology becomes both clear and acute. If the exponential advancement highlighted by Kurzweil continues apace, as seems likely though not certain, then the outcome is highly unpredictable. It could be bliss for all, or unspeakable destruction — or something inbetween. We could all wind up dead — killed by software, wetware or nanoware bugs, or other unforeseen phenomena. If humanity does vanish, it could be replaced by radically more intelligent entities (thus satisfying de Garis's Cosmist aesthetic) – but this isn't guaranteed; there's also the possibility that things go awry in a manner annihilating all life and intelligence on Earth and leaving no path for its resurrection or replacement.

To make the dilemma more palpable, think about what a few hundred brilliant, disaffected young nerds with scientific training could do, if they teamed up with terrorists who wanted to bring down modern civilization and commit mass murders. It's not obvious why such an alliance would arise, but nor is it beyond the pale. Think about what such an alliance could do now — and what it could do in a couple decades from now, assuming Kurzweilian exponential advance. One expects this theme to be explored richly in science fiction novels and cinema in coming years.

But how can we decrease these risks? It's fun to muse about designing a 'Friendly AI' à la Yudkowsky, that is guaranteed (or near-guaranteed) to maintain a friendly ethical system as it self-modifies and self-improves itself to massively superhuman intelligence. Such an AI system, if it existed, could bring about a full-on Singularity in a way that would respect human values — i.e. the best of both worlds, satisfying all but the most extreme of both the Cosmists and the Terrans. But the catch is, nobody has any idea how to do such a thing, and it seems well beyond the scope of current or near-future science and engineering.

Realistically, we can't stop technology from developing; and we can't control its risks very well, as it develops. And daydreams aside, we don't know how to create a massively superhuman supertechnology that will solve all our problems in a universally satisfying way.

This train of thought leads naturally to the possibility of creating what I've called an AI Nanny' — a *mildly* superhuman supertechnology, whose job it is to protect us from ourselves and our technology — not forever, but just for a while, while we work on the hard problem of creating a Friendly Singularity.

## The 'AI Nanny'

More specifically, what I mean by an 'AI Nanny' is an advanced Artificial General Intelligence (AGI) software program with

- General intelligence somewhat above the human level, but not too dramatically so — maybe, qualitatively speaking, as far above humans as humans are above apes
- Interconnection to powerful worldwide surveillance systems, online and in the physical world
- Control of a massive contingent of robots (e.g. service robots, teacher robots, etc.) and connectivity to the world's home and building automation systems, robot factories, self-driving cars, and so on and so forth
- A cognitive architecture featuring an explicit set of goals, and an action selection system that causes it to choose those actions that it rationally calculates will best help it achieve those goals
- A set of preprogrammed goals including the following aspects:
  - A strong inhibition against modifying its preprogrammed goals
  - A strong inhibition against rapidly modifying its general intelligence
  - A mandate to cede control of the world to a more intelligent AI within N years (where N could be, say, 10, 100 or 5000)
  - A mandate to help abolish human disease, involuntary human death, and the practical scarcity of common humanly-useful resources like food, water, housing, computers, etc.
  - A mandate to prevent the development of technologies that would threaten its ability to carry out its other goals
  - A strong inhibition against carrying out actions with a result that a strong majority of humans would oppose, if they knew about the action in advance
  - A mandate to be open-minded toward suggestions by intelligent, thoughtful humans about the possibility that it may be misinterpreting its initial, preprogrammed goals

Obviously, this sketch of the AI Nanny concept is highly simplified and idealized — a real-world AI Nanny would have all sort of properties not described here, and might be missing some of the above features, substituting them with other related things. My point here is not to sketch a specific design or requirements specification for an AI Nanny, but rather to indicate a fairly general class of systems that humanity might build.

The nanny metaphor is chosen carefully. A nanny watches over children while they grow up, and then goes away. Similarly, the AI Nanny would not be intended to rule humanity on a permanent basis – only to provide protection and oversight while we ‘grow up’ collectively; to give us a little breathing room so we can figure out how best to create a desirable sort of Singularity.

When I first reflected on this idea, my personal reaction was to find it rather odious. But after further reflection my view is more ambivalent. One point I considered is that, in spite of a personal streak toward rule-breaking, I’m not a political anarchist — because I have a strong suspicion that if governments were removed, the world would become a lot worse off, dominated by gangs of armed thugs imposing even less pleasant forms of control than those exercised by the US Army and the CCP and so forth. I suspect government could be done a lot better than any country currently does it — but I don’t doubt the need for some kind of government, given the realities of human nature. It may be that the need for an AI Nanny falls into the same broad category. It seems possible that, like government, an AI Nanny is a relatively offensive thing, that is nonetheless a practical necessity due to the unsavory aspects of human nature.

We didn’t need government during the Stone Age — because there weren’t that many of us, and we didn’t have so many dangerous technologies. But we need government now. Fortunately, these same technologies that necessitated government, also provided the means for government to operate.

Somewhat similarly, we haven’t needed an AI Nanny so far, because we haven’t had sufficiently powerful and destructive technologies. And now, these same technologies that *may* necessitate the creation of an AI Nanny, also may provide the means of creating it.

### The Argument for Building an AI Nanny

To recap and summarize, a plausible ethical argument for trying to build an AI Nanny would be that:

1. It’s impracticable to halt the exponential advancement of technology (even if one wanted to)

2. As technology advances, it becomes possible for individuals or groups to wreak greater and greater damage using less and less intelligence and resources

3. As technology advances, humans will more and more acutely lack the capability to monitor global technology

development and forestall radically dangerous technology-enabled events

4. Creating an AI Nanny is a significantly less difficult technological problem than creating an AI or other technology with a predictably high probability of launching a full-scale positive Singularity

5. Imposing a permanent or very long term constraint on the development of new technologies is undesirable

It would be interesting and valuable to run through this argument with the analytical detail of Chalmers' article on the Singularity; but this is merely a brief commentary, so a rough summary will have to do for now.

The fifth and final premise is normative; the others are empirical. None of the empirical premises are certain, but all seem likely to me. The first three premises are strongly implied by recent social and technological trends. The fourth premise seems commonsensical based on current science, mathematics and engineering.

These premises lead to the conclusion that trying to build an AI Nanny is probably a good idea. The actual plausibility of building an AI Nanny is a different matter – I believe it is plausible, but of course, opinions on the plausibility of building any kind of AGI system in the relatively near future vary all over the map.

The above argument is interesting from two points. First, it might be correct. And second, even if it is incorrect for some reason, it is possible that if sufficiently powerful organizations come to believe it, they may create an AI Nanny of some form anyway.

## Complaints and Responses

I have discussed the AI Nanny idea with a variety of people over the last year or so, and have heard an abundance of different complaints about it — but none have struck me as compelling. Here follows a partial, roughly-sketched list of counterarguments and my counter-counterarguments.

*It's impossible to build an AI Nanny; the AI R&D is too hard.* — But is it really? It's almost surely impossible to build and install an AI Nanny this year; but as a professional AI researcher, I believe such a thing is well within the realm of possibility. I think we could have one in a couple decades if we really put our collective minds to it. It would involve a host of coordinated research breakthroughs, and a lot of large-scale software and hardware engineering, but nothing implausible according to current science and engineering. We did amazing things in the Manhattan Project because we wanted to win a war—how hard are we willing to try when our overall future is at stake?

It may be worth dissecting this 'hard R&D' complaint into two sub-complaints:

- *AGI is hard*: building an AGI system with slightly greater than human level intelligence is too hard (i.e. in Chalmers' terms, AI+ is too hard)
- *Nannifying an AGI is hard*: given a slightly superhuman AGI system, turning it into an AI Nanny is too hard (i.e. in Chalmers' terms, the particular kind of AI+ that is an AI Nanny is too hard)

Obviously both of these are contentious issues.

Regarding the 'AGI is hard' complaint, at the AGI-09 artificial intelligence research conference, an expert-assessment survey was done (Baum *et al.*, 2011), suggesting that a least a nontrivial plurality of professional AI researchers believes that human-level AGI is possible within the next few decades, and that slightly-superhuman AGI will follow shortly after that.

Regarding the 'Nannifying an AGI is hard' complaint, I think its validity depends on the AGI architecture in question. If one is talking about an integrative, cognitive-science-based, explicitly goal-oriented AGI system like, say, OpenCog (Goertzel, 2009) or MicroPsi (Bach, 2009) or LIDA (Friedlander & Franklin, 2008) then this is probably not too much of an issue, as these architectures are fairly flexible and incorporate explicitly articulated goals. If one is talking about, say, an AGI built via closely emulating human brain architecture, in which the designers have relatively weak understanding of the AGI system's representations and dynamics, then the 'nannification is hard' problem might be more serious. My own research intuition is that an integrative, cognitive-science-based, explicitly goal-oriented system is likely to be the path via which advanced AGI first arises; this is the path my own work is following.

*It's impossible to build an AI Nanny; the surveillance technology is too hard to implement.* — But is it really? Surveillance tech is advancing very rapidly, for reasons more prosaic than the potential development of an AI Nanny. David Brin's book *The Transparent Society* (1999) gives a rather compelling argument that before too long, we'll all be able to see everything everyone else is doing.

*Setting up an AI Nanny, in practice, would require a world government.* — This seems patially valid. It would require either a proactive assertion of power by some particular party, creating and installing an AI Nanny without asking everybody else's permission; or else a degree of cooperation between the world's most powerful governments, beyond what we see today. Either route seems conceivable. Regarding the second cooperative path, it's worth observing that the world is clearly moving in the direction of greater international unity, albeit in fits and starts. Once the profound risks posed by advancing technology become more apparent to the world's leaders, the required sort of international cooperation will probably be a lot easier to come by. Hugo de Garis's most recent book *Multis and Monos* (2010) riffs extensively on the theme of emerging world government.

*Building an AI Nanny is harder than building a self-modifying, self-improving AGI that will retain its Friendly goals even as it self-modifies.* — I find this rather implausible. Maintenance of goals under radical self-modification and self-improvement seems to pose some very thorny philosophical and technical problem — and once these are solved (to the extent that they're even solvable) *then* one will have a host of currently-unforeseeable engineering problems to consider. Furthermore there is a huge, almost surely irreducible uncertainty in creating something massively more intelligent than oneself. Whereas creating an AI Nanny is 'merely' a very difficult, very large scale science and engineering problem.

*If someone creates a new technology smarter than the AI Nanny, how will the AI Nanny recognize this and be able to nip it in the bud?* — Remember, the hypothesis is that the AI Nanny is significantly smarter than people. Imagine a friendly, highly intelligent person monitoring and supervising the creative projects of a room full of chimps or 'intellectually challenged' individuals.

*Why would the AI Nanny want to retain its initially pre-programmed goals, instead of modifying them to suit itself better? — for instance, why wouldn't it simply adopt the goal of becoming an all-powerful dictator and exploiting us for its own ends?* — But why *would* it change its goals? What forces would cause it to become selfish, greedy, etc? Let's not anthropomorphize. 'Power corrupts, and absolute power corrupts absolutely' is a statement about human psychology, not a general law of intelligent systems. Human beings are not architected as rational, goal-oriented systems, even though some of us aspire to be such systems and make some progress toward behaving in this manner. If an AI system is created with an architecture inclining it to pursue certain goals, there's no reason why it would automatically be inclined to modify these goals.

*But how can you specify the AI Nanny's goals precisely? You can't right? And if you specify them imprecisely, how do you know it won't eventually come to interpret them in some way that goes against your original intention? And then if you want to tweak its goals, because you realize you made a mistake, it won't let you, right?* — This is a tough problem, without a perfect solution. But remember, one of its goals is to be open-minded about the possibility that it's misinterpreting its goals. Indeed, one can't rule out the possibility that it will misinterpret this meta-goal and then, in reality, closed-mindedly interpret its other goals in an incorrect way. The AI Nanny would not be a risk-free endeavor, and it would be important to get a feel for its realities before giving it too much power. But again, the question is not whether it's an absolutely safe and positive project – but rather, whether it's better than the alternatives!

*What about Steve Omohundro's* Basic AI Drives *(2008)? Didn't Omohundro prove that any AI system would seek resources and power just like human beings?* — While the arguments in this paper are powerful, they are mainly evolutionary in nature. They apply most plainly to the case of an AI competing against other roughly equally intelligent and powerful systems for survival. The posited AI Nanny would be smarter and more powerful than any human, and would have, as part of its goal content, the maintenance of this situation for N years. Unless someone managed to sneak past its defenses and create competitively powerful and smart AI systems, or it encountered alien minds, the premises of Omohundro's arguments don't apply.

*What happens after the N years is up?* — This is unclear, which is part of the argument for creating an AI Nanny — the point is that after N years of research and development under the protection of the AI Nanny, we would have a lot better idea of what's possible and what isn't than any of us do right now.

*What happens if the N years pass and none of the hard problems are solved, and we still don't know how to launch a full-on Singularity in a sufficiently reliably positive way?* — One obvious possibility is to launch the AI Nanny again for a couple hundred more years. Or maybe to launch it again with a different, more sophisticated condition for ceding control (in the case that it, or humans, conceive some such condition during the N years).

*What if we figure out how to create a Friendly self-improving massively superhuman AGI only 20 years after the initiation of the AI Nanny — then we'd have to wait another N-20 years for the real Singularity to begin!* — That's true of course, but if the AI Nanny is working well, then we're not going to die in the interim, and we'll be conducting an enjoyable existence while we wait.

*But how can you trust anyone to build the AI Nanny? Won't they secretly put in an override telling the AI Nanny to obey them, but nobody else?* — That's possible, but developing the AI Nanny via an open, international, democratic community and process would diminish the odds of this sort of problem happening.

*What if, shortly after initiating the AI Nanny, some human sees some fatal flaw in the AI Nanny approach, which we don't see now. Then we'd be unable to undo our mistake.*— True, that is a risk of the approach.

## Conclusion

Chalmers opines that, of all the possible defeaters that might prevent a Singularity from occurring in the next few centuries, the most likely are the 'motivational defeaters' — i.e. factors that might cause humanity or its AI+ creations to *intentionally* avoid launching a rapidly exploding process of intelligence that creates smarter intelligence that creates smarter intelligence….

An argument against the likelihood of motivational defeaters would be that, even if some humans and AI+'s lack the motivation to create AI++, probably other humans or AI+'s will not suffer this lack, and will go ahead and create AI++ anyway. General technological advance seems likely to progressively decrease the amount of computational and other resources to create more and more powerful AI systems, making it likely that eventually a relatively small group of enthusiasts could launch AI++ regardless of the opinions of others.

However, one way that a motivational defeater could prevail, in spite of the factors mentioned in the above paragraph, would be via the creation of an 'AI Nanny' – an AI+ system with ample physical empowerment and the explicit goal of preventing the occurrence of AI++, either permanently, for a certain fixed period of time, or else till certain pre-specified criteria are met. Whether the creation of an AI

Nanny is plausible — or will become plausible prior to the creation of AI++ — is certainly an open question. However, I have argued that the AI Nanny route may become very appealing to many parties sometime before the Singularity occurs, due to its potential for circumventing some of the dangerous ethical dilemmas the Singularity presents.

Specifically, humanity is facing a situation where increasing danger (potentially even human extinction) is wreakable by small groups of disaffected individuals; and the possibility of AI++ is also fraught with danger and uncertainty, because of our lack of any rigorous scientific theory of AI++ systems. The creation of an AI Nanny has potential to dampen the risks of terrorism and runaway AI++, and giving humanity and AI+ a bit of breathing space to figure out if AI++ is the right thing to do, and if so to do it right. Whether or not creating an AI Nanny is the best course for humanity, it could be the course taken anyway, if sufficiently powerful organizations decide it is the best option.

## References

Bach, J. (2009) *Principles of Synthetic Intelligence*, Cambridge: Cambridge University Press.

Baum, S.D., Goertzel, B. & Goertzel, T.G. (2011) How long until human-level AI? Results from an expert assessment, *Technological Forecasting & Social Change*, **78** (1), pp. 185–195.

Bostrom, N. (2002) *Journal of Evolution and Technology*, **9** (1).

Brin, D. (1999) *The Transparent Society*, New York: Basic Books.

Chalmers, D.J. (2010) The singularity: A philosophical analysis, *Journal of Consciousness Studies*, **17** (9–10), pp. 7–65.

de Garis, H. (2004) *The Artilect War*, Pittsburgh, PA: Etc Press.

de Garis, H. (2010) *Multis and Monos*, Pittsburgh, PA: Etc Press.

de Garis, H. (2011) *Merge or Purge*, [Online], http://hplusmagazine.com/2011/05/19/merge-or-purge/

Friedlander, D. & Franklin, S. (2008) LIDA and a Theory of Mind, in Goertzel, B. & Wang, P. (eds.) *Artificial General Intelligence (AGI-08)*, Memphis, TN: IOS Press.

Goertzel, B. (2009) OpenCogPrime: A Cognitive Synergy Based Architecture for General Intelligence, *International Conference on Cognitive Informatics*, Hong Kong.

Joy, B. (2000) Why the future doesn't need us, *Wired Magazine*, (April).

Kurzweil, R. (2005) *The Singularity Is Near*, New York: Viking.

McKibben, B. (2004) *Enough*, New York: St. Martin's Griffin.

More, M. (2009) *Singularity and Surge Scenarios*, [Online], http://strategicphilosophy.blogspot.com/2009/06/how-fast-will-future-arrive-how-will.html

Omohundro, S. (2008) The basic AI drives, in Wang, P., Goertzel, B. & Franklin, S. (eds.) *Proceedings of the First AGI Conference*, vol. 171, Frontiers in Artificial Intelligence and Applications, Memphis, TN: IOS Press.

Wallach, W. & Allan, C. (2008) *Moral Machines*, Oxford: Oxford University Press.

Yudkowsky, E. (2002) *Creating Friendly AI*, http://singinst.org/upload/CFAI.html

Yudkowsky, E. (2004) *Coherent Extrapolated Volition*, http://singinst.org/upload/CEV.html

Yudkowsky, E. (2009) *The Challenge of Friendly AI*, [Online], http://itc.conversationsnetwork.org/shows/detail3387.html