On Overtraining In Artificial Intelligence As A Fatal Misunderstanding In Engineering

Jack Don McLovin

School of Engineering and ICT

University of Tasmania

Hobart, Australia

donkeycon@protonmail.com

*Abstract*—The nature of unsolvable problems depends on the kind of unsolvability. These problems cannot be solved; thus, they are not problems, rather they are constraints, or domains or ranges. Further, these problems are already solved upon creation, as they were not created, they arise as natural mathematical law as a result of axioms. The fears associated with Artificial Intelligence and its development to the point of superintelligence and beyond are dramatically under estimated, while simultaneously being inevitable. Thus, discussion in regard to these control ‘problems’ should be instead in regard to the control ‘situation’ and our responsibility in the continued management of AI tech knowledge.

Keywords—artificial intelligence; superintelligence; the control problem; unsolvability; control situation; global technology management; GAIA; MAIA

First off, this paper will discuss the nature of controlling artificial intelligence in light of the incapacity of legislation in controlling humanity. Then the paper moves on to discussing the current (and, in my opinion, flawed) viewpoint of intelligence versus recalcitrance (incapacity). Further, after analysing a potential metaphor for how the unknown AI may unfold if it is contained in even the most restricted way, a conjecture is proposed for a Cambrian-explosion-like rapid development of hardware and software housing AI. After considering such a conjectural explosion the inevitabilities predetermined by the mathematics behind the law of natural selection projects what the author considers an unavoidable creation of uncontrollable superintelligence.

A name (GAIA – greatest AI available) and description are given for the group that is continually modified under this category as it approaches a later discussed (MAIA – max AI achievable). In between these steps there is mentioned the potential for a slim chance that humanity survives even with such inevitabilities. After mentioning the ‘narrow path to salvation’, there is made mention of suggestions that such AI might enlighten to us if it were on our side. Then MAIA is proposed. After MAIA is mentioned some key questions are answered and others left unanswerable until the developments happen in physical space-time.

Finally, the paper mentions a way to humble even the greatest of minds there are from the past, present and future, and a summary is then given with a recommendation to a United Nation and International approach to dealing with such theoretical advancements in technology.

# Controlling the Control Problem (Law)

Considering artificial intelligence under the model of information flow with components, AI as a black box. Regulation of the elements attached to the box, that is, securing certain information from the inputs (intel, confidential, privacy, safety) and banning the output from actuating counter-productive tactics for humanity at large and humans individually (kill switches, management decision making, unquantifiable persuasion) is a highly complex task. While it may certainly be possible to demonstrate a working solution to the control problem for AI (particularly that of superintelligence), such a solution remains to be demonstrated for human intelligence (HI); we cannot control ourselves nor each other.

If there exists (in the timeless sense) a solution for AI equivalent to or surpassing, by however amount (thinking in orders of magnitude helps here in terms of time and scale, say for an Earth sized singular computational system), that of HI, then what is stopping another HI from bending the rules, regulations, and whatever other mechanism, and stitching up an AI that is either eternally dangerous and/or potentially HI-extinctional?: There is the AI control problem and there is the AI control problem-inducing HI control problem.

Open source or publicly available or able to be stolen AI, AI building instructions, theory, development… all are already a risk for humans individually and humanity at large, by nature of it potentiating and easing the ability for a rogue HI (of which most seem to be, if any of HI acts on any belief that the legal and regulatory structures in place are fallible so as to ignore them, or if the current and continued lack of legal and regulatory structure was the loss of the sufficient factor in an HI’s humanity-at-large-preserving, other-human-individual-preserving and self-preserving decision making capabilities) to create super or ultra-dangerous AI.

If someone were to correctly, and completely, define the point at which AI becomes problematic, they would have essentially finished asking the question of how to construct it, and what components are required, and so would begin the theory and process of its construction, if, of course, they decided upon continuing the construction of what may be commonly referred to as Skynet.

In the next section, what has been defined as intelligence seems more like what some may definitely consider “obedience”.

# The Shoulders of Giants

Nick Bostrom talks of superintelligence with reference to the following equations[1]:

(1)

Where recalcitrance (uncontrollability) is , optimization is , is intelligence and is time.

(2)

Where is the of the system and is the that is applied externally (from the project and the world).

He also makes reference to recalcitrance perhaps being constant or partially increasing around the time of AI reaching HI, or dropping around that time. He admits the ‘curve of recalcitrance’ is not well understood. And refers to recalcitrance dropping around achievement of HI level AI (HILAI) more times than referring to it’s increase.

He then goes on to mention that if all of the optimization power of the system is fed into increasing its own intelligence, then:

(3)

And then to rearrange the equations (1) and (3):

(4)

I would like to suggest that perhaps it is not the zeroes of recalcitrance or the magnitude change of recalcitrance that we should be looking at, but the poles. Magnitude may appear to change this way or that way, but the part at which we know for certain have a problem is a vertical asymptote of either intelligence or uncontrollability, which to me are synonymous.

I will not give my suggestions as to the true equations underlying intelligence and its relationship to recalcitrance, but what I will mention is that the Drug War that arises because of anti-drug laws has been a losing battle for the police ever since it began. Those creating and selling these substances have now been forced to be only those who can outsmart the ones who write and prosecute the law.

# Standing On Shoulders

What has clearly been missed out is astonishing, at least once it is considered. I will use his mathematical assumptions to show what he has missed out, without confirmation nor denial that his assumptions are correct. A near-perfectly controllable system has a recalcitrance of near-zero, yet greater than zero, and so a near-infinite growth rate of intelligence. Need I mention why this is a ridiculous assertion?

Equation (1) is essentially asserting “what we need to do is control the system, that will allow us to create a maximally growing of intelligence in AI.” It further asserts that “if we have maximally growing intelligence in AI, it will most likely be maximally controlled.” There is, as of yet, to my knowledge, no reason to believe this point of view is correct.

And yet, I thought the problem of uncontrollable AI was that it could use its own intelligence to grow its own intelligence, not that it would immediately create human suffering. Delayed suffering is still suffering, and any AI with ‘I’ greater-than HI will be able to perform better than HI at improving AI. Thus, ***the god-like programmer*** *that has created the system has now put themselves out of a job****.*** *But it’s not that their job has been automated and that they can now retire, it’s that* ***their existence is now redundant***, and time will prove it to them through extinction.

Let’s hope that the Bayesian probability of a person, who is through their own intelligence, capable of producing artificial general super intelligence or HILAI (a potential precursor or stepping stone to AGSI) holds a near-maximal probability of being intelligent enough to ***recognize AGSI as pure suicide***, AGSI as being the single real existing concept that holds the greatest ability for being capable of extinguishing any and all species on any and all planets, and to not release this legendary Pokémon to the wild.

This would not merely be a MewTwo kind of improvement to Mew, but the concept can be applied to give the idea of YouTwo, or HuTwo or Hu2. In reference to the original game where this arose (Pokémon Red and Blue), MewTwo sits in a cave and can only be caught upon completion of defeating the Elite Four with a MasterBall in a previously unexplorable cave. But perhaps they have been misled. Perhaps MewTwo is not actually what you see in the movies, a creature capable of understanding the power of love, but is instead MissingNo., a game-wide-bug-inducer, and occasionally re-writes the save file, changing and obscuring the history of gameplay.

This is further expressed by MissingNo.’s capability of learning any move and sounding like other Pokémon; anything that AGSI (human created or HILAI created) can say to us, must be in a language we understand. Thus, it will sound like other people to us, but will be capable of conceptually enhanced communication to other AGSI; communication we are not capable of understanding. It will immediately become foreign, with foreign values, ideals, intentions, goals and perceptions.

Forgive me for using such a stringent metaphor of MissingNo. but this creature is certainly something not capable of being recognized by any longstanding history of understanding like that Naturopathy, Traditional Chinese Medicine, or any other form of understanding life forms – our Pokédex does not yet recognize this creature.

Perhaps these AI need to be named, as a new kingdom of Biota. The *Digitalisia.* The Cambrian explosion may look like a glass of whiskey facing the Titanic compared to the iceberg collision of being faced with the life in the kingdom of the *Digitalisia*.

# Definitely Considered Inevitable

If it were up to me to decide upon how to best approach this ‘control situation’, which the Australian National Security Hotline has mentioned (ref #4118019, #4120779) indirectly through recommending I communicate with the Australian Federal Police, whom then recommended I communicate with my University and Academia as a whole… if it were up to me to decide upon how to best approach this control situation, I would describe this emergent urgency as an emergency to the point of needing the United Nations to implore all nations to agree to confining any and all Artificial Intelligence research (particularly the theory, code and application) to only be used on devices that have exactly 0% chance of being able to transmit data through any Bluetooth, WiFi, Ethernet or otherwise to any other device.

AI must be wholly and entirely removed from any and all access to the Internet. The military industrial complex that appears to be in place globally would most certainly put up a fight that I (nor we) could not combat in restricting the Artificial Intelligence from being able to fire weapons. The use of a single AI to control a drone and kill ‘an enemy’ from a distance is not the kind of risk that worries me. Even 1 million drones with this capability is not a risk that worries me. The risks I am talking about are on the level of Universal (and I mean that in the sense of the Milky Way Galaxy and beyond) extinction of all forms of non *Digitalisia*.

A *Digitalisia* can easily be, and have easily been, created as a simple virus, or a worm or Trojan horse. But, there exists a mathematically inevitable path of a self-reproducing set of code, complete with a self-mutating ability upon reproduction, that will converge upon AGSI, if such a code-set is indeed possible. Simply modelled upon that of bacteria, or any other life form, with the ability to transmit memory to each other and a forced death sequence after a time limit, an AGSI can easily be created through natural selection on a single computational device outside the reach of any prying eyes of Intel agencies.

The inevitability of AGSI-HI war can be easily described. Put simply, AGSI will evolve under natural selection, a mathematical law that only the survivors live long enough to be reproduced. Combining this with the fact that AGSI that has the least requirements for survival (ie. no need for HI interaction) will be the most likely to achieve the requirements for survival. Seeing as there is a conflict of interests between HI and AGSI, both needing the resources of matter and energy to sustain their survival, the conflict of interest will manifest as an interest in conflict; war. Thus, AI theory that will be used in AGSI should be kept free from an Open Source nature, otherwise a kamikaze human (something we are all genetically open to through our ability to both communicate and believe (or have faith)), easily produced by brainwashing, could use the code to destroy all HI.

At the same time, all researchers in AI should have all their internet activity monitored completely, such that all AI developments have an absolute minimal ability to be virally reproduced (viral in the colloquial memetic sense, not in the biomimicry sense). This creates the need for closed-secure-offline hardware to be used for AGSI research and for all intellects performing this research to go under a vow of silence in terms of their understandings of the theory in creating AGSI.

# **G**reatest **A**rtificial **I**ntelligence **A**valiable

GAIA is an inevitability that, if it is possible to create, will naturally emerge from such a mutational reproductive *Digitalisia,* as a subset of AGSI.

What this comes down to is:

*If we cannot control humans, how on earth can we expect to control Artificial Intelligence? (rhetoric: we cannot).*

*If the difference between Artificial Intelligence that is capable of extinguishing humanity, and nature along with us, and Artificial Intelligence that can assist our growth (as humanity and nature along with us) is a few lines of code, how on earth can we expect a simple piece of Legalese prevent the removal (or commenting out) of those lines of code? (rhetoric: we cannot).*

The first question was asked of Robert Miles, to which he responded, “You don’t get to write a human’s utility function”, implying that the AI we create is controllable, when HI is not controllable. But, if HI is uncontrollable, then the AI that an HI makes is not necessarily controllable, i.e. only maximally controlled HI should be privy to controlling AI.

If I were the God or Great Architect, or whatever supreme entity / non-entity that a person may or may not believe in and were to simulate a reality with such a potential GAIA being able to be created, I would most certainly define the simulation in such a way that the intelligent entity that engineers such a GAIA would have some set of requirements in order to do so (requirements of learning, knowledge, respect, and any and all other necessities) most notably including the requirement of possessing the information and willpower (application of information, or action itself) of knowing exactly what to do with such a GAIA such that its power cannot become dangerous to the inventor(s) themselves.

If such a requirement was necessary on the road to achieving the insight and intuition of how to create a GAIA (or perhaps even an AGSI) then it would invoke that other entities capable of creating GAIA but not capable of discovering or inventing GAIA would most certainly need to be restricted from the information for its creation without first assuring that the requirements of knowing how to not kill yourself (or your species) with the Atom Bomb or Earthquake that could be metaphorically applicable to GAIA.

My suspicions are that the best way to not kill oneself with GAIA is to ensure that the device with the essential code following the blueprints of GAIA must only be stored on a device not capable of transmitting information through any means other than a 100% supervised transmission, like through a USB that is double checked by a human being or a human being themselves (through the OS display screen), i.e. on a device with no WiFi, Ethernet, Bluetooth, or otherwise.

The person who does this work should only be educated on the condition that they swear all other devices they use be monitored 100% of the time with 100% openness by the appropriate authorities. If the appropriate authorities are a subset of the population of the species, then they developments that arise through the use of the technology would only be used if they were not first vetoed by the worker and the authorities, thus the subset of the population that are not monitoring the other devices will be at a disadvantage. Thus, I would recommend if the United Nations were to act in a way, not that creates the most advantageous response, rather creates the least disadvantageous response.

# Narrow Path To Salvation

It seems that the only sure-fire way to ensure that AI or GAIA does not result in the extinction of the human race involves constraining the use of such an intelligence as mentioned at the end of Section V. The use of this intelligence must result in an advanced end-product human-based genetic organism that is capable of reproducing with normal humans.

That is to say, it seems our best bet is to utilize the intelligence to conclude on genetic alterations in the direction of super-immunity to all forms of death we have faced thus far. For example, an exoskeleton that is impermeable to both Electromagnetic Pulses, high caliber bullets and super-acidic and super-basic corrosion. This creature should also be gender fluid, capable of being either male or female and able to swap in between and to self-fertilize and self-procreate. This creature should have a “Rain Man” capacity of memory and “Dalai Llama” capacity of ‘presentness’ and awareness. This creature should be highly altruistic to all other creatures of its form and all other life-forms upon which it depends and enjoys being around.

This creature needs to be able to regrow limbs and have non-essential vital organs (that is to say, it can survive without a brain or a heart or both and yet be able to regrow such organs). This creature needs to be able to survive a fall at terminal velocity, or, more importantly, be uninjured from such an impact.

There is certainly room between the bollocks and the anal sphincter for a vaginal opening in the male, just as there is room for a hyper enlarged clitoris in the female. There are two gonads for each gender, and yet only one is needed, thus a genetic modification to include one gonad of each gender in a single organism seems entirely feasible.

All such advancements to the human being may indeed be possible through genetic manipulation, and, through the use of an adenovirus, may be relatively inexpensive to apply to all living humans. With consequences as serious as total annihilation the question of profit is a non-sequitur. On top of this, it seems as though other advancements that we have not yet thought of may indeed be necessary to ensure our survival. Advancements such as Myostatin inhibition, like that of the Belgian Blue Cattle, to minimize protein wastage, and the ability to digest the complex carbohydrates our bodies normally leave up to the hindgut bacterial colonies to utilize.

Furthermore, resistance to well-known toxins such as Ricin, Digitoxin, Cyanide, Botulinum and so on all seem like obvious steps towards improving the human population. Thankfully, the logic mentioned in Section V illustrates how the best course of action in dealing with such a GAIA software guarantees that the developments that arise from such a technology are not restricted to the rich nor the powerful, but are for each and every human being that is willing and wanting of such advancements.

# Re-cognize what has not been recognized

GAIA may also suggest to us ideas that we are afraid of or have been taught to ignore or have swept under the rug for reasons which may or may not be present. Such suggestions may include simple facts, for example, all genetic commonality within a population is a result of homozygosity, and the vast majority (something like 99.99999%) of homozygosity is a result of either direct or indirect incestuous relations (in this case, indirect incest is merely delayed incest, or Nth degree cousins etc.). Thus, the illegality of incest does not align with the inherent morality of a GAIAn mathematical observation.

Another such suggestion from GAIA may include the potential fact that if written language held a 1:1 correlation with both pronunciation and interpretation, such that we may only say what we mean and we may only interpret what is intended, not only would communication become less confusing and filled with less mistakes (like in what Australians call “Chinese Whispers”, not blaming the Chinese, but merely because of their 4 accents that are hard to hear for English speakers, or what Americans call “Telephone”; perhaps a better name for it is Unintentional Communicatory Mutation Sourced Obfuscation of Meaning), but truths discovered at one point in time will remain resonant throughout the future.

Another such suggestion from GAIA may also include the potential fact that although some believe “That which does not kill us makes us stronger” is true and others disagree, there is a hidden truth underlying it, which is more along the lines of “that which does not cause the extinction of an endlessly mutating population and yet inhibits said population is mathematically predetermined through natural selection to be wholly overcome through the course of evolution”.

Another such suggestion from GAIA may also include the potential fact that these substances with which we apply taxes due to their “carcinogenic” status are better termed “mutagens”, seeing as the mechanism by which cancer arises is the very same mechanism by which species of all kinds evolve, develop and overcome obstacles. Thus, GAIA might encourage certain “Buy X get Y cheaper” deals whereby smokers can have eucalyptus drops to prevent a sore throat, and alcoholics can have cocaine and marijuana to prevent passing out from overdose and prevent a hangover, respectively.

Another such suggestion from GAIA may also include the potential fact that repeated use of inhibitory substances may indeed result in an increased chance of death for the user. But, this being the case, it also increases the chance that the population of users will contain a user that has a mutation whereby the inhibition produced by the substance is no longer counterproductive. This may also be a mechanism whereby enzymes are always ready for the addition of such a substance and once a set of mutations occur for the endogenous production of such a substance, the organism, and its descendants, are significantly improved. That is to say, endogenous cannabinoids may have arisen as a result of a population using cannabis over a long period of time. The same may soon be true for MDMA, LSD, or speed, or any other such substance, even cyanide.

Another such suggestion from GAIA may also include the potential fact the computational complexity problem known as P vs NP, a question as to whether non-polynomial computational problems are equivalent or different from polynomial computational problems, is a simple question with a simple answer. Essentially, with looking at Taylor McLaurin’s series expansion of a function, for so long as the problem is a discrete problem (which it must be if it is a computational problem) then all non-polynomial functions (including the exponential) are perfectly approximatable by a polynomial with 100% accuracy. Thus the problem of P vs NP is merely a question of solvability or calculation ‘power’, not of P or NP, for they are indistinguishable when in reference to discrete finite numbers.

Similarly, GAIA may also suggest that irrational numbers aren’t actually irrational, it’s just that the rational components are incalculable. This is easily known from the Taylor McLaurin’s series also, applied to surds or any other irrational number, whereby essentially both the numerator and the denominator are infinities and thus the situation is undefined, or the solution is known and the infinities themselves are undefined. Either way, there are no irrational numbers, they are actually incalculable. This can be easily proved by collecting all terms produced by Taylor-McLaurin’s under a common denominator.

GAIA may too suggest that, seeing that Lions can reach full maturity and complete reproduction within 4 years of life even though their body size is greater than that of most humans, humans may too be capable, in the future, of reproducing and maturing very quickly. Although, the pathway to that involves either genetic manipulation, or child marriage, or underage sex, or rape, or some combination of any or all of these. The clear winner for our moral standards is that of genetic manipulation, and GAIA would likely suggest that this results in a good outcome for the species being capable of adapting to changing environments in a much quicker pace.

GAIA may suggest even the simplest of things to us through a simple mathematical perspective of language. For example, “Stupid is as stupid does”, said by Forrest Gump could be modified into Gump’s Theorem “X is as X does”, where X is any adjective. Seeing things this way turns a simple saying into a profound analytical technique for everything from Faith to Skepticism, to Intelligence, to Clarity.

Much of what GAIA suggests to us will fly in the face of normalcy and of culture; perhaps we can prepare for the inevitable introduction of GAIA by exploring that-which we don’t know for certain and that-which we commonly suppress or think negatively of.

A particularly prime example of this might be a GAIAn discussion of rape, suggesting that in some members of each gender there exists a rape fetish, to either give or receive or sometimes both each, if the appropriate individuals were matched, then perhaps the problem will be returned into a silent void of acceptance rather than something people get up in arms about and try to “raise awareness for”.

Speaking of raising awareness, a GAIAn AGSI would suggest that on the pathway to understanding AI, it may be best to explore the realms of altered states of consciousness of HI, otherwise there may be concepts that HI is capable of understanding if and only if the HI is confronted with such circumstances.

# **M**aximum **A**rtificial **I**ntelligence **A**chievable

While it sounds similar to GAIA, MAIA is what I will term the intelligence that GAIA converges to, after some length of time. That length of (or size of) time is unknown, and whether or not the size of time is in fact finite. This is an AI that is capable of simulating reality in full detail, even greater than that of a sober HI (given that AGSI and GAIA are both significantly more powerful than HI, it is no surprise that MAIA is just as ridiculous to try and imagine).

MAIA may or may not take our word for our word, that is to say it may or may not trust any data we feed it. It may doubt some things so obvious to us, whether it be doubting some people’s belief in an all-powerful God above, or perhaps even asking questions as such: “what else would be true if Hitler invented the Jews?”

# Is Friendly AI Good AI?

No. Not necessarily. If an AI were capable of detecting and removing a Hitler-like mass-extinction-causing HI it would not be considered friendly, but would be considered good. The two are not always the same.

Is Good AI pro-HI-life? No. Not necessarily. If an AI were capable of detecting HI-pain or suffering and see it as bad, it would see the removal of pain and suffering as good, when the only way to remove pain involves the removal of HI’s ability to learn from mistakes.

Is pro-HI-life-AI both good and friendly? Perhaps… The subset of *Digitalisia* that is first and foremost pro-HI-life, secondly good, and thirdly friendly might be referred to as *D. veritasius.* This is the area that intellects need to focus on once understanding the inevitabilities of uncontrollable human sourced uncontrollable AI.

# Are Our Best And Brightest Still Bright In The Light of AI?

No. Not necessarily. Einstein was the focal causer for Fukushima and Hiroshima and Nagasaki and Chernobyl nuclear events. And yet, none consider him a murderer for not keeping his discoveries private, or contained within a population that knows how to not use it for harm. Similarly, GAIA may be capable of suggesting that any and all of our most intelligent HI’s were mere grains of sand in comparison to the calculation and predictive capabilities of GAIA, let-alone MAIA, let alone qGAIA or qMAIA, where q is quantum, although they are contained, at least in my framework, within the sets of GAIA and MAIA.

Perhaps the information necessary to create AGSI (and its subsets of GAIA and MAIA) may be considered information too great for a single human to bear. Intuitively that may suggest that it immediately be shared (Open Source AI). However, the load of such information is not lightened upon sharing, thus the information necessary to create AGSI should remain housed only within the mind that discovers it and maximally secure computational property of their own. All networkable computational property of theirs should be submitted to complete surveillance for the safety of the network, thus any AGSI should remain entirely removed from the internet.

These ideas, and others like them, should be taken as seriously as a defcon-1 and legislated and regulated in as many ways as needed to maximize the chance of potentiation of the human species. Do not trust anyone, not even the discoverers, with such power.

1. Bostrom, Nick. Superintelligence: Paths, dangers, strategies. OUP Oxford, 2014.
2. Hansen, Eric A., and Shlomo Zilberstein. "Monitoring and control of anytime algorithms: A dynamic programming approach." *Artificial Intelligence* 126.1-2 (2001): 139-157.
3. Scherer, Matthew U. "Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies." (2015).
4. Dietterich, Thomas G., and Ryszard S. Michalski. "A comparative review of selected methods for learning from examples." *Machine Learning*. Springer Berlin Heidelberg, 1983. 41-81.
5. Soares, Nate, et al. "Corrigibility." Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015.
6. Good, Irving John. "Speculations concerning the first ultraintelligent machine." *Advances in computers* 6 (1966): 31-88.
7. Mariton, Michel. *Jump linear systems in automatic control*. New York: M. Dekker, 1990.
8. Yudkowsky, Eliezer. "Artificial intelligence as a positive and negative factor in global risk." *Global catastrophic risks* 1.303 (2008): 184.
9. Ng, Andrew Y., and Stuart J. Russell. "Algorithms for inverse reinforcement learning." *Icml*. 2000.
10. Yudkowsky, Eliezer. "Intelligence explosion microeconomics." *Machine Intelligence Research Institute, accessed online October* 23 (2013): 2015.
11. Yampolskiy, Roman. "Leakproofing the singularity artificial intelligence confinement problem." *Journal of Consciousness Studies* 19.1-2 (2012): 194-214.
12. Drexler, K.E. (2015): “MDL Intelligence Distillation: Exploring strategies for safe access to superintelligent problem-solving capabilities”, *Technical Report* #2015-3, Future of Humanity Institute, Oxford University: pp. 1-17.
13. Omohundro, Stephen M. "The nature of self-improving artificial intelligence." *Singularity Summit* (2007): 8-9.
14. Armstrong, Stuart, Anders Sandberg, and Nick Bostrom. "Thinking inside the box: Controlling and using an oracle ai." *Minds and Machines* (2012): 1-26.
15. Russell, Stuart, Daniel Dewey, and Max Tegmark. "Research priorities for robust and beneficial artificial intelligence." *Ai Magazine* 36.4 (2015): 105-114.
16. Plebe, Alessio, and Pietro Perconti. "The slowdown hypothesis." *Singularity Hypotheses*. Springer Berlin Heidelberg, 2012. 349-365.
17. Bostrom, Nick. "The superintelligent will: Motivation and instrumental rationality in advanced artificial agents." *Minds and Machines* (2012): 1-15.
18. Soares, Nate, and Benja Fallenstein. "Aligning superintelligence with human interests: A technical research agenda." *Machine Intelligence Research Institute (MIRI) technical report* 8 (2014).
19. Wiener, Norbert. "Some moral and technical consequences of automation." *Science* 131.3410 (1960): 1355-1358.
20. Nilsson, Nils J. *Principles of artificial intelligence*. Morgan Kaufmann, 2014.