Can we Build a Superintelligence Without Being Superintelligent?

Robert Sternhell

October 9, 2017

A thesis submitted in partial fulfilment of the requirements for the degree of Bachelor of Arts (Honours) in Philosophy, University of Sydney, October 2017.

Word Count: 14897

Acknowledgements

I would like to express my gratitude to my supervisor, Professor David Braddon-Mitchell, for his help with my thesis. He has also been a role model for me throughout my undergraduate, from my first philosophy course to my final year. His humour and enthusiasm showed me how enjoyable philosophy could be and I hope I can do the same for others.

I would like to extend my thanks to all my lecturers and tutors. I could not have developed any skill at this without constant challenges and discussion.

I would also like to thank my friends and family for supporting and encouraging me for the past five years. Their discussions, questions and challenges have helped me clarify nebulous ideas into more intelligible ones.

Lastly, special thanks should be given to those who proofread my drafts. Their feedback proved extremely beneficial to this paper.

Table of Contents

Abstract

If we create an entity of greater intelligence to us, a superintelligence, it has the possibility to explode in intelligence, creating more and more intelligent entities. If the intelligence explosion argument holds, then the most important step to developing a powerful superintelligence is the development of the first superintelligence. This paper presents objections to the possibility of humans developing this first superintelligence. I argue that this is because we lack required knowledge about them, due to our epistemic position of not being superintelligent. I engage primarily with arguments from David Chalmers and Nick Bostrom about what superintelligences are and the nature of the intelligence explosion. I add my own detail to these areas and explore how to increase intelligence. I argue that my objections stem from flawed expectations of superintelligence such that we ought to change them. I then present my own alternative expectations for superintelligence.

Keywords: *artificial intelligence, AI, The Singularity, intelligence, intelligence explosion, superintelligence*

1 Introduction

In the near future, we may face the development of machines that exceed our intelligence in every way. Ray Kurzweil claims that this will happen as early as 2039.[1] As proponents of the singularity have become more popular, philosophical inquiry into the nature of these beings has increased. These intelligences were first defined by I.J Good as 'ultraintelligences' which "*were machines that can far surpass all the intellectual activities of any man however clever*".[2] The term superintelligence has since become popular and has been defined by Nick Bostrom similarly as "*an intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills*."[3] We have already developed artificial intelligences (AIs) that individually surpass humans in many domains. AI researchers are continually making progress in automation of increasingly complicated tasks. However, superintelligences are more than a toolbox of AIs. They are self-contained intelligences that use their intelligence to complete, or at least attempt, any kind of goal in the same way that we can.

If a superintelligence can do this, then we ought to be able to task them with constructing superior AI and they should be able to use their skills to do so. If successful, the next generation could do the same and so on. Good referred to this process as the '*intelligence explosion*'.[4] This also known as 'The Singularity' (although

[1] Ray Kurzweil, *The Singularity is Near: When Humans Transcend Biology* (New York: Viking, 2005).
[2] Irving John Good, "Speculations Concerning the First Ultraintelligent Machine", *Advances in Computers* 6 (1966): 33.
[3] Nick Bostrom, "How Long Before Superintelligence?", *International Journal of Future Studies* 2 (1998): 1.
[4] Good, "Speculations": 33.

it has fallen out of favour in academic circles recently).[5] If uninterrupted, the cycle continues towards ever greater, or potentially maximum, intelligence because each iteration is able to build something more intelligent than itself. With greater intelligence, each generation not only overcomes previously insurmountable development obstacles but also ought to be much cleverer than us in other sorts of intelligent tasks. Proponents of the intelligence explosion believe that the final, or later, superintelligence will be extremely powerful because of its intelligence. Anything that intelligence can bring us, superintelligences will bring it faster and better. It will be able to counter our every opposition and best us with its superior intellect. This power has the potential to destroy humanity or launch it into salvation, according to its will.

Bostrom extensively discusses what the will and capacity of a superintelligence will be.[6] He begins with superintelligences possessing goals, and his analysis of its will and nature flows from the assumption that it will do whatever is best to complete its goals. This seems plausible enough as intelligence ought to correspond with capacity to reason and therefore a superintelligence will act as a perfect reasoner. He then concludes that they will prioritise outcomes like increasing intelligence and self-preservation over all else, because they are beneficial for goal completion. Humanity would be vulnerable to incidental annihilation by an AI if we became an obstacle to completion of its goals. Bostrom argues that for most goal systems, humanity is either an obstacle or a potential resource.

[5] Vincent C. Müller, "Risks of General Artificial Intelligence", *Journal of Experimental & Theoretical Artificial Intelligence* 26, no. 3 (2014): 297-301.
[6] Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2013), 127-140.

What is striking about these intelligences is how alien they are to us. This is in part because they are emotionless and possess narrow priorities but their reasoning is even more mysterious. It is difficult to imagine exactly what a perfect reasoner does to solve any problem in the real world. If I asked a perfect reasoner to fetch me a cup of tea, it is not obvious exactly what it would do to process that information or how I would receive my tea (if at all). Our first superintelligence will not be such a thing. It will be an imperfect reasoner, relatively similar to ourselves. For this reason, it is difficult to ascertain which problems it will excel in and which it will not. However, it is important that this intelligence has the necessary abilities to iterate on itself, or the intelligence explosion will not occur. To overcome this, superintelligences could possess general intelligence. This kind of intelligence can bridge all necessary domains, or is able to properly enlist domain specific intelligences. This allows successive superintelligences to iterate on themselves as each version ensures a greater capacity for more AI improvement.

This is a high burden on the first superintelligence we create. For it to begin the intelligence explosion, it must solve whatever problems occur in the creation of that intelligence such that the next intelligence is also a general intelligence capable of this. If this is true then there are two possible conclusions. First, we are general intelligences (or the collaboration of humans equate to one) capable of building superior intelligences such that we can begin this cycle. Second, we are not general intelligences or otherwise cannot begin the cycle. I will investigate the first conclusion by exploring whether we can build the first superintelligence. I argue that we may be unable to do this because of the limitations our intelligence imposes on the design of superintelligences. Therein, I will argue that the second possibility is

more likely. In doing so, I will explore what intelligence, and subsequently superintelligence, is and how we ought to define improvements in either. I will suggest that our definition of superintelligence is at fault and that we ought to change it accordingly.

This paper will be in six parts. Section **2** will explore what intelligence is. Section **3** will define superintelligence and general intelligence. In section **4**, I will argue that there is a difference between quality and speed superintelligences. In section **5**, I will investigate whether we can design a general intelligence or improve upon human intelligence. Section **6** will assess methods of identifying superintelligences. In section **7**, I will attempt to use my previous analysis to diagnose these problems and suggest alternative definitions for superintelligence.

2 What is intelligence?

The term intelligence has often referred to features unique to the human experience like consciousness, self-awareness or ethical personhood. At the same time, we often refer to non-human creatures as possessing some form of intelligence. Non-biological entities even have intelligence and we label them as artificial *intelligences*. For this paper, features like consciousness, qualia or advanced thought will not be necessary conditions of intelligence.

One could argue that this account of intelligence is too functionalist. John Searle distinguishes between 'weak' and 'strong' AI to address a similar problem.[7] Strong AI

[7] John R. Searle, "Minds, Brains, and Programs", *The Behavioral and Brain Sciences* 3 (1980): 417-457.

possess consciousness, experience and mental states. Whereas, weak AI can produce intelligent outputs but lack any of these qualities. Let us consider the set of potential functionally superintelligent beings. Each program is at least a weak AI and some of them are strong AIs as they also have mental states. If intelligences require strong AI characteristics to be superintelligent, then the set of superintelligent beings is entirely comprised of strong AIs, each of which are also functionally superintelligent and will not be left out of my analysis. However, if there are some functional superintelligences that are only weak AIs, then they are left out of any analysis that presumes mental states. By including both weak and strong AI, I ensure that we can cover all potential superintelligences. After all, intelligent labour is not restricted to entities possessing these features, so I will not restrict my analysis to them.

Nick Bostrom argues that intelligence is an information processing algorithm that receives information from the world and produces an output.[8] Therefore, anything that has intelligence, has a certain capacity to process information to solve a specific problem. It does not need to be an organism or even an agent of any kind, it must only be something that performs this optimisation task.

In other words, an intelligence is something that receives a problem and produces some kind of solution to it. In the vast majority of observed cases of organic intelligences, the problem is in the form of external stimuli and the solution is a produced response to that stimuli.[9] By this definition, intelligence is abundant in the world. A gazelle has intelligence in avoiding lions and mosquitos are intelligent in

[8] Bostrom, *Superintelligence*, 26-29.
[9] A lack of response altogether is also a form of a response and is part of information processing algorithms although is less interesting for this analysis.

avoiding squashings from predatory palms. These kinds of intelligences are domain specific in that they only apply to one particular set of problems (or even exactly one problem). By this account, many modern computer programs, and animals, not only have intelligence but have exceeded human intelligence in certain domains.[10]

Alan Turing defined intelligence in a similar fashion with the Turing Test.[11] The test would assess whether someone could differentiate between a human replying through a text console and a computer producing replies over the course of a half hour conversation. The limitation of this test is that we cannot generalise from the scope of the test to all future possible cases. Therefore, passing a Turing Test cannot be a sufficient condition for intelligence although it is a necessary condition. It is worth noting that Turing himself did not consider this test a sufficient condition for intelligence, narrowly or broadly. He saw it as a measure of the capacity of that machine to produce outputs we associate with intelligence within the domain of text-based conversation.[12]

Any functionalist measurement of intelligence will share this limitation with the Turing Test. However, functional intelligence is exactly what we need to perform intelligent labour. Therefore, we can measure intelligence (at least the kind that we need) by measuring how good the outputs of a system are. For example, a Chess computer is more intelligent than me in the domain of Chess, if it is more likely to win

[10] This is certainly an expansive definition of intelligence. However, we sometimes refer to abstract phenomena like evolution as intelligences; or refer to the intelligent optimisation of river systems; or pathfinding of plants and fungi.
[11] A. M. Turing, "Computing Machinery and Intelligence", *Mind* 236 (1950): 433-460.
[12] Steven Harnad, "Minds, Machines and Turing: The Indistinguishability of Indistinguishables", *Journal of Logic, Language, And Information* 9, no. 4 (2000): 9; Noam Chomsky, "Turing on the "Imitation Game"", in *Parsing the Turing Test* (Dordrecht: Springer, 2009).

similarly difficult games of Chess than I am.[13] More intelligent Chess computers will optimise better with respect to the game state and perform moves that win more games. This measure of intelligence does not infer anything about the way the algorithm produces the result.[14] Rather, intelligence refers to the capacity of a system to produce intelligent outcomes within the domain, or domains, of which we are measuring.

There are two ways intelligence can be increased. We can improve quality or speed.[15] A faster intelligence is one that able to produce an answer more quickly. Whereas a greater quality intelligence must be able to produce a more *intelligent* answer. This means that if it has unlimited time to calculate an answer, it will eventually terminate (stop calculating and return an answer) with an answer that better solves the problem. For example, if we compare two Chess algorithms based on their quality. The better-quality Chess intelligence, if unrestricted by time, is more likely to win because it will terminate with better moves on average.

There are many ways that we could potentially increase quality intelligence by having better problem-solving methods. In everyday language, we often provide long and detailed stories of how we solve problems. One need only a cursory glance at

[13] There is a metagame aspect to this. One could imagine that a chess champion, Paul, could be better than all other players except one player, Fred, who through use of some special strategy can beat Paul. However, Fred is unable to beat all other players. One could argue that Fred is better than Paul because he can win that game, or that he is worse than Paul because he is less likely to win games in against peers. This seems to me less a problem with the measuring of intelligence and more a problem for what we consider the goal of chess. Is it to win more often, or to win against specific peers within a certain metagame?

[14] By this definition, even something like Ned Block's 'Blockhead' would be intelligent. I do not wish this to mean that it has any of the qualities that Block was arguing against (and labelling them with the term intelligence), rather that it is able to perform the job of an intelligence. See Block (1981).

[15] Bostrom introduces a similar distinction about superintelligences. He separates them into speed, quality and collective intelligences. I will discuss collective and speed intelligences under the umbrella of speed intelligences and quality intelligences as is. See Bostrom (2013).

business and self-help literature to find enormously detailed descriptions of how to solve problems that allege to promote creativity and effectiveness. However, this does not help us find out why any problem-solving method is better than another. Let us consider the following examples. Suppose I wish to buy ice cream and I have a choice between chocolate and vanilla. In my decision, I might refer to comparative qualities of each my options, like flavour and cost. I process the details of each option and apply my beliefs, like which flavour of ice cream I tend to enjoy, as a method of assessing their merits and deficiencies. I can assign a value (likely implicitly) to the total merits of vanilla and chocolate. If I choose the vanilla ice cream it is because I have concluded that vanilla has a greater total value than chocolate.[16]

This is one example of how we might solve problems but it is obviously not the process of all intelligences. For example, to solve the problem of '65537 + 1729', I do not go through every possible answer, number or otherwise. I examine the problem, recognise what it is asking me for and enlist the necessary algorithm. For a sum, we require the 'addition' algorithm and performing all the steps (like carrying the one) until the problem is solved. If we want to know how to increase the quality of intelligence, I will need to provide a good way to compare these seemingly disparate methods and improve them.

Let us consider a simple case of domain specific intelligence, a Tic Tac Toe solver. The algorithm can only receive inputs in the form of a state of the board and whether it is noughts or crosses to move. All other inputs are nonsense indiscernible from no input at all. Any modern processor can easily calculate every possible board

[16] There are many other ways to understand decisions like this. I have chosen this one to illustrate one way, not necessarily the correct way.

state in the game, known as the search space, and identify which moves only result in a draw or victory. This strategy is called 'brute force' as it solves the problem by looking at every possible option through sheer magnitude of computation. This is a maximum quality intelligence because it will always produce the best possible move. In principle, brute force can be adapted to any kind of problem such that, if run to termination, it will produce the best possible answer.

We can obviously play games like Tic Tac Toe but we also seem to be able to play games like Chess and Go, of which the search space is approximately 10^{50} and 10^{170} possible legal board states respectively.[17] It is easy enough for our brains to look at any given end state and determine whether it is a victory or a loss, based on the rules of the game. However, it takes a prohibitively long amount of time, say until the heat death of the universe, to brute force these problems. Instead, we utilise heuristics that filter the search space, through the structure of the algorithm we use, until we are left with one option (or sometimes more than one).

Let us now return to the addition case. Implicitly, the addition algorithm dismisses many answers, ranging from numbers that are too small or too large to responses that are not even numbers. It does not explicitly consider whether '2 + 2' could equal 'Q' or '−410'. However, if we were to know nothing about '2 + 2', then an answer like 'Q' or '−410' or 'Tutu' could all be possible answers, each equally likely. The additional algorithm latches on to the presentation of a problem in the form of '2 + 2' and performs a structural process that implicitly rejects answers like 'Q' or 'Tutu'

[17] Claude E. Shannon, "Programming a Computer for Playing Chess", *Philosophical Magazine* 41, no. 304 (1950); Victor Allis, *Searching for Solutions in Games and Artificial Intelligence* (Maastricht: Rijksuniversiteit Limburg, 1994).

because there is no way for it to produce those answers. It limits the search space immediately to only numbers and then systematically performing operations until it terminates with one answer. When performed correctly, is a maximum quality intelligence because it will always terminate with the best answer.

Intelligences can also be maximum quality without sorting the information in ways as elegant as this. An intelligence for answering '2 + 2', that always responded with '4' would produce the right answer as often as the addition algorithm.[18] Although it is not 'adding' in the same sense as the other, it is a maximum quality algorithm because in all circumstances in which it is used, it produces the best answer. This might seem like an absurd conclusion as surely such a system is not performing the work of an intelligence. However, we cannot know from the outside whether something is using a system like this or truly calculating something.[19] This is enough for me to want to include these in our analysis such that we include all functionally intelligent systems.

The search space reduction is also inconspicuous in the ice cream case but still occurs. To start, my ice cream algorithm ignores huge chunks of the search space that includes options like wild screaming and not buying ice cream, or controlled screaming and not buying ice cream. Moreover, the algorithm will structurally weigh up qualities like my beliefs about ice cream but structurally ignores other beliefs like

[18] Research suggests that we do in fact answer '2 + 2' in a way that is like this algorithm rather than actually 'adding' anything. See Kahneman (2011).

[19] We might also express doubt as to whether we can know exactly which algorithm is being used based on what outputs are produced. This problem is best expressed in Kripke (1982) with respect to whether we can know if an algorithm that someone else uses is truly doing what it says it is doing. I do not need to claim anything about how something actually processes information. I only claim about what it can functionally produce so this issue is not particularly important for my argument.

the political situation in Mongolia, or the date of the First Norman Conquest. An algorithm of lower quality might completely ignore my beliefs about ice cream and choose chocolate, leaving me disappointed. An even lower quality algorithm might ignore all possibilities that involve buying ice cream, leaving me without an ice cream and more disappointed. Both are less intelligent because they fail to select the correct option, purchasing vanilla ice cream, as compared to an algorithm that does. Maybe, if the goal of the algorithm is keeping me healthy, a refusal to buy ice cream is good but if the task is to select between vanilla and chocolate ice cream based on what I want to eat, it fails miserably.

These cases show us that we can improve the quality of an intelligence by improving how it filters the search space. There are two ways to do this. We can better assess options it has access to or increase the area of the search space it can access. Neither option is sufficient for more intelligent outcomes. If an algorithm is already finding the best solution out of what it has access to, then the former method will do no good. If the solution it currently produces is better than anything outside the accessible search space, then the latter does no good.

Increasing the quality of intelligence will generally increase the amount of calculations required. However, often the priority of an intelligent system is to terminate quicker. There are two ways that something can terminate more quickly. First, the algorithm can terminate in less steps, which tends to reduce quality but not necessarily. Second, the algorithm can be identical but be carried out more quickly through greater computational power.

The first method provides a powerful evolutionary advantage. A gazelle that calculates every possible path to avoid a lion has a maximum quality lion-avoiding intelligence. However, our maximum quality genius gazelle is a dead gazelle. Organisms are often better off having an imprecise algorithm that makes quicker decisions and requires less resource hungry brainpower. There are some maximum quality algorithms that bear this price but there are also some maximum quality algorithms in which the gazelle always runs to the left and is always optimal. There are also algorithms in which the gazelle always runs to the right and always suboptimal but it is certainly better than the gazelle always dying. Either algorithm makes trade-offs in quality, although not always, in exchange for expediency.

It is less obvious that the second method, of increasing computational power, leads to greater intelligence. Suppose I want to run a complicated hurricane simulation. I could choose to use a supercomputer or my laptop. The supercomputer would perform the simulation hundreds, if not thousands, of times faster than my laptop. However, it would be essentially performing the *same* task. If I ran it on my laptop long enough, it would terminate with an identical result (assuming there were no random or pseudo-random elements to the simulation and it does not require some minimum amount of memory). It is not obvious whether we would consider the supercomputer's performance more intelligent but it is certainly more useful and produces more favourable outcomes. Regardless, it is not important that we accept this as increase in intelligence, only that it is different to an increase in quality of intelligence.

To improve the speed of an intelligence, all we need to do is plug in more processing power, or make more of the same mechanical process.[20] However, an intelligence with greater speed will never give us better answers. Although it is often important to receive answers quickly, only increases in quality can get us better solutions. Therefore, the distinction between quality and speed is that increased speed can never give us a better answer, only the same answer more quickly. This is useful for some domains, less so for others. I will return to this distinction later when I compare speed superintelligences to quality superintelligences.

3 What is a superintelligence?

Thus far, I have only explained how something can be intelligent in a single domain. Consider modern Chess algorithms I have already mentioned. The best Chess program on a modern supercomputer is not just better than the average human player, or the best human player, it is likely better than any foreseeable human player (bar cognitive enhancement).[21] Lucky for us, a Chess intelligence has only limited impacts on the real world. The world is not comprised of Chess pieces on a Chessboard. The rules and systems of the world are vastly more complicated and numerous. A Chess intelligence can only ever eclipse us within the domain of the rules of Chess and is therefore benign. However, domain specific intelligences are not necessarily benign. One can certainly cause serious good or ill with domain specific intelligences. However, intelligences that are not limited by domains can potentially perform any task to this level, rather than only one.

[20] Maybe it is not always *that* easy to expand capacity but the process is, at least in principle, clear.
[21] Monty Newborn, *Beyond Deep Blue* (London: Springer London, 2011), 257-262.

There are two possible ways that something can be intelligent in all domains. First, it can possess a collection of domain specific intelligences that are able to span all problems. It is possible that there are an infinite number of domains and if so, we would therefore require an infinite number of intelligences. However, Bostrom negates this problem by defining superintelligences as being superior in human domains so non-human domains are irrelevant to him.[22] This seems reasonable enough as it is unlikely that humans possess intelligence in non-human domains and we probably do not interact with non-human domains anyway. The problem for this interpretation of general intelligence is that our intelligence is not a cluster of domain specific intelligences. A new kind of problem does not baffle us because we lack specific intelligence. Rather, we are able to parse out novel problems and develop solutions to them. It is certainly possible to design an intelligence like this but it will be insufficient for problems in domains it is unfamiliar with. This is a big problem for an intelligence that will need to confront likely novel problems in an intelligence explosion.

The second way is to have intelligence that is not domain-limited, often labelled as 'general intelligence'. Bostrom argues that superintelligences are general intelligences and that generality is essential.[23] General intelligence is difficult to conceptualise but we can begin with the claim that we possess it. David Chalmers and Bostrom both claim this.[24] We could assume that humans alone possess it but this seems implausible as animals display a great deal of ostensibly advanced human behaviours. As do young children who are bested by animal counterparts in other

[22] Bostrom, *Superintelligence*, 68.
[23] Bostrom, *Superintelligence*, 26-27.
[24] David Chalmers, "The Singularity a Philosophical Analysis", *Journal of Consciousness Studies* 17, no. 9-10 (2010): 16-17; Bostrom, *Superintelligence*, 27.

skillsets, including problem solving. Therefore, it is more plausible that we are on a continuum of intelligence, such that some level is possessed by dogs, birds, apes and worms but to differing degrees in each.[25] Each animal can be placed on a curve of general intelligence with humans at some point above the other animals. Amongst humans, geniuses may only be slightly further along than an average person. It is difficult to determine what the distances between each animal on the scale ought to be. We might place worms far behind us but our primate relatives, dolphins, octopuses and other intelligent animals very close to us on this curve.[26] Kurzweil positions animals based on number of neurons they possess but this can be misleading as humans do not have the largest brains but do seem to be the most intelligent.[27]

A superintelligence would be slightly further along us on this curve. This means that is smarter than the smartest human, rather than humanity as a whole. Bostrom points out that it is difficult to determine exactly how intelligent humanity is. He argues that humanity could be a superintelligence, a collective superintelligence, if it had sophisticated enough communication methods but it is not at the present time.[28] A superintelligence would be smarter in all domains than the best humans because of higher general intelligence. If the analogy holds true then the character of such an intelligence becomes apparent. A relatively small increase in intelligence to our

[25] Maybe not *all* organisms but certainly the more intelligent ones like apes. Although, if we want to claim this, it is difficult to place an exact threshold.
[26] Bostrom, *Superintelligence*, 64.
[27] Kurzweil, The Singularity, 59-75.
[28] Bostrom notes that global intelligence has risen a thousand times over since the Pleistocene and therefore would be a superintelligence compared to those individuals. He does not believe that possessing sporadic intelligence makes something a collective intelligence, rather it is the extent to which cooperative systems allow for quick and effective exchange of information. He does not believe that the scientific community, for example, is a superintelligence. Regardless it seems plausible that we could increase our communication structures until it meets this burden. See Bostrom (1998) & Bostrom (2013).

closest relative, the Bonobo, has allowed us to utterly dominate them. To them, most of our world is likely to be unfathomable, unpredictable and insurmountable. For this reason, Good commented that such intelligences would be humanity's last invention.[29]

However, this analogy does not really tell us what it means to be further along the curve. So, why approach the problem with an analogy like this? The primary reason is that the intelligence explosion requires an intelligence that spans all necessary domains. Each subsequent superintelligence must iterate on the previous level to make a superior intelligence. Suppose that a problem requires intelligence X_n, where X is the domain and n is the level of intelligence, to solve. The next increase may require an n level intelligence in any domain A, B, C… to continue iterating. This requires that either all domains follow from an increase in some overall intelligence, like general intelligence, or that an increase in any domain specific intelligence necessarily gives the system the ability to increase another area of intelligence. The problem with the latter possibility is that it does not preclude an explosion from getting stuck improving only certain parts. Suppose that intelligence A_4 can improve intelligence B_4 to B_5 which can improve only A_4 to A_5 and so forth. It is not clear why this must extend to any other area of intelligence or why it ought to continue. To ensure that an explosion does not get stuck, or stray off course to an unbalanced intelligence (potentially dangerously), we must have general intelligence, of which A...Z intelligences develop from it, and contribute to it.

[29] Good, "Speculations", 31-88.

Chalmers defends this point by distinguishing only between cognitive capacities and self-amplifying capacities.[30] He claims that improvement capacities will surely explode because they will continue to self-improve and that cognitive capacities will correlate with them. If so, then there ought to be a similar resulting increase in cognitive capacities, or at least some increase. I agree with Chalmers that self-improvement capacities will likely correlate strongly with at least some cognitive capacities. If it has general intelligence, then improvements to that algorithm necessitate resulting increases in cognitive capacity, as long as the starting intelligence is general enough to produce a true general intelligence. If superior intelligence is a mixture of many kinds of intelligences interacting, then the problem is not about whether the correlation exists but whether expertise in a given domain helps solve problems in the next important domain and does not get stuck.

What remains to be answered is what it takes to create a greater general intelligence. There are two possible options. We can create intelligences like us in a greater magnitude, or with greater quality. In the next section, I will assess the difference between the two and argue that quality superintelligence is required for the intelligence explosion and the character of superintelligence described above.

4 Is there a difference between speed and quality superintelligences?

The intelligence explosion sets a high burden for quality superintelligence. There is an alternative kind of thing we might call a superintelligence, which Bostrom

[30] David Chalmers, "The Singularity: A Reply to Commentators". *Journal of Consciousness Studies* 19, no. 7-8 (2012): 148-150; Chalmers, “Philosophical Analysis”, 16-19.

labels a speed superintelligence, that would comprise of a greater magnitude of human intelligences rather than improvements to the quality of intelligence. To examine speed intelligences let us begin with a single human intelligence running at high speeds. There are many ways to do this. Bostrom argues that 'Whole Brain Emulation' is a promising method for mass production of human level AI.[31] Chalmers is sympathetic to this process but argues that we are more likely create a human level AI in the near future through conventional design.[32] I am not particularly motivated towards either position. Both systems should be similar anyway as we will likely base human level AI, implicitly or otherwise, on human intelligence. However, part of the argument for a general superintelligence is that we possess general intelligence. If we begin with Whole Brain Emulation we can at least be confident that our starting point is a general intelligence rather than an intelligence of ostensibly equal quality and generality but unknown underlying design. This will also help us better conceptualise the limitations on a speed superintelligence.

The emulation process involves scanning a single human brain and replicating it using computer hardware. An average processor can perform billions of operations *per second*. Although it is difficult, and often misleading, to assign an exact number of operations per second to a human brain, it is safe to say it is considerably less than a modern processor.[33] Human brains are heavily parallelized compared to computers but we can feasibly imagine that the human brain could be spread across multiple processing units (as modern supercomputers tend to do anyway) and hence we could

[31] Nick Bostrom and Anders Sandberg. "Converging Cognitive Enhancements". *Annals of the New York Academy of Sciences* 1093, no. 1 (2006): 201-227.
[32] Chalmers, "Philosophical Analysis", 29.
[33] Nick Bostrom and Anders Sandberg, "Whole Brain Emulation: A Roadmap", *Technical Report #2008-3*, Future of Humanity Institute, Oxford University (2008).

replicate some of this parallelisation. This results in a single human intelligence running thousands of times faster than normal. Were we to upload the mind of Albert Einstein, for example, we would reap the results of lifetimes of thought in mere months, or days or even minutes depending on how much hardware we utilize. To see the limitations of Einstein 2.0, consider the following case.

> To our dismay we discover that a human brain is far too complicated for us to scan using current hardware. All is not lost as we find that we have precisely enough hardware to support uploading of the brain of a dog. Lucky for us, like a human, the speed of a dog's brain can be greatly increased using modern processors. We place our trusty dog Snuffles in the scanning device, create a digital duplicate and run his little brain at one million times speed. In just seven minutes he has lived thirteen years and in a short few months he will have lived thousands of lifetimes in which to contemplate our problems.

Even though Snuffles has vastly more processing power than any individual human however smart, it seems like Snuffles will not provide us with particularly meaningful answers regardless of how long we run the program. This is because Snuffles suffers from two problems. The exclusion problem and the false positive problem.

4.1 The exclusion problem

The exclusion problem is that we have no guarantee that the best solution, or even a good one, will necessarily occur to any level of intelligence. As I previously argued, any kind of intelligence explicitly or implicitly avoids most of the search space. Given that we are not fundamentally modifying the intelligence algorithm of the subject (say Snuffles), we have no reason to expect that it will change its exclusion patterns. Snuffles excludes most of the good solutions that we would like

his superspeed dog brain to explore because they are not options that his brain would have ever explored. In the same way that of all the dogs on earth, none of them have started to quote Shakespeare or recommended nuanced policy advice, it seems that we ought not to expect Snuffles to either. Increasing the processing speed of Snuffles does not fix this problem in any meaningful way. Without improving the quality of his intelligence, he will never be able to explore more of the search space than a dog does.

This kind of problem equally applies to Einstein 2.0. It seems to me that when we talk of genius, Einstein's or otherwise, intellect it is not a claim about how many more mental hours they have spent, it is a claim about producing insights that their counterparts were not able to. Einstein did not ponder over every possible arrangement of characters, or possible laws of physics, in order to discover his field equations. Rather, he possessed a particular method of problem solving that systematically narrowed the search space resulting in novel and ground-breaking insights. Speeding up this process does not change what options he explicitly or implicitly assesses, it only changes how fast it can be done. It may be possible that a human, given an enormous amount of time, could assess all possible options. However, this is equivalent to humans brute forcing the problem. I have already mentioned problems with this strategy. Employing humans to do it only increases the computational burden.

4.2 The false positive problem

Even if we ignore the exclusion problem, our superspeed intelligence is useless if it does not terminate on a better answer than a regular speed human. Suppose we

were to replace Einstein 2.0 with someone who is inept at mathematics. They could be very smart but simply lack this aptitude. If we were to pose them a mathematics question, at some point they must stop thinking and return an answer. Why ought we expect this answer to always be correct? It seems to me that humans often make mistakes and our mathematically inept person ought to as well. We could give them as much time as they wanted and it seems like they still might fail to answer questions correctly. Moreover, our mistakes are not only in instances where we have yet to find the right answer, we often think we have found the right answer. We maintain that we are correct until someone who knows better has explained it to us or some external system has validated the answer.

For some kinds of questions, we can validate answers easily but for other questions it is less clear what that validation might be. In the case of Snuffles, he will terminate exactly as he would without super speed, by asking to go for a walk or play fetch. The problem here is twofold. First, the problem-solving algorithms we use are structured, both during our lifetimes and evolutionarily, to favour termination rather than the best answer. Any amount of superspeed does not change that fact. Second, without a fundamental improvement to whatever system we use to assess whether an answer is correct, a human brain on superspeed will often reach the same conclusion as one not on superspeed because both believe they are correct. It may be possible that the additional time allows a single human to think through a problem that requires more than a single lifetime to ever consider and cannot be shared amongst multiple individuals. I am not sure which problems, if any, fit this category. Regardless, it seems difficult to argue that *all* or even many problems are of this kind.

The exclusion problem and the false positive problem ought to tell us that a human brain run at super speed suffers from systematic problems that will often lead to answers very similar to humans run at normal speeds. Therefore, a single human on superspeed has the same kind of limitations that the dog on superspeed does. In the same way that Snuffles could not compete with any human thinker, it seems that a superspeed human ought to have the same trouble competing against something that has a higher quality intelligence.

Let us now consider a cluster of human intelligences running at superspeed, rather than just a single human intelligence. Although it would be considerably easier to create a million duplicates of Einstein 2.0, they are likely to all agree with each other as they are fundamentally identical systems. I will assume that we can create variation in duplicated minds in the same way we have significant variation in thinkers today.[34] Our AI will have Einsteins, John von Neumanns, David Lewises, etc. The important difference between this system and the previous, is the capacity for the minds to have differing approaches to problems. Thus, the system starts to more closely resemble the way that progress in ideas actually occurs in the real world.

There are two ways that we might facilitate this interaction. First, it could have some kind of cognitive link that allows the brains to function as one. It is unclear what this might look like and would require a more detailed understanding of intelligence than we have. Second, it could communicate much in the same way that we do. This seems like a much more achievable method of communication that ought

[34] This could be done by scanning as many geniuses and experts as we have today. However, it may be very difficult to create novel variation in intelligence from a single scanned brain without an in depth understanding of human intelligence.

to be simple enough once we are already capable of mass brain emulation. This would essentially amount to a colossal superspeed academic conference of the greatest minds to work on the world's problems.

This would certainly be impactful. To say otherwise would be to deny that academic conferences are useful at all.[35] In the same way that we ought to be concerned if an enemy nation had access to a thousand more scientists than we previously believed, we ought to be concerned with the development of speed superintelligence, like a superspeed conference. However, the system still suffers from the same objections as Einstein 2.0. It lessens the exclusion problem by assessing more options because each mind is likely to, by virtue of variation, explore different areas of the search space. Similarly, it alleviates some of the false positive problem by providing many different validation methods from different minds. However, it eliminates neither of these problems and is therefore subject, at least to some extent, to them. Moreover, it inherits its own class of problems by being a communicative system. Sometimes groups get the wrong answer, not because the right answer was not suggested but because the structure of discussion wrongly eliminated the right answer.

Nonetheless, it could replicate the output of thousands of years of academic conferences in a fraction of the time. This seems like it would be hugely intelligent. After all, this is how we have progressed in the past few centuries, which has ushered in the fastest acceleration in knowledge and technology in our history. However, the least intelligent superintelligence (only slightly qualitatively smarter than us) does

[35] Although some may argue this is a tenable position.

not need to be smarter than the superspeed conference for the difference to matter. There are two reasons for this.

First, if we create one slightly qualitatively smarter intelligence then we can eventually grant it the same benefits of scale. In the same way that a superspeed conference full of genius level intelligences is much better than a superspeed conference full of individuals of average intelligence, we ought to think the higher quality superspeed conference would be smarter. Superior intelligences may also be able to communicate in more effective ways, increasing this difference. We often make a similar claim about more intelligent humans, that they are able to communicate better.[36] Moreover, if we return to Snuffles once more and suppose we create a superspeed conference of mentally varied dogs, we will probably find the system to be no better than our original superspeed Snuffles. This is, at least in part, because dogs lack certain features of intelligence that allow for complicated and effective communication such that they cannot collaborate to spread intellectual labour. It is possible that such a difference could exist between us and superintelligences as well.

Second, if we accept that an intelligence can create something that is qualitatively cleverer than it is, then we ought to accept that the greater intelligence can make an intelligence that is even greater still. This is accepting that the intelligence explosion holds. If this is true, then a higher quality intelligence that has this capacity, entails a much greater intelligence shortly after.

[36] There are some exceptions. One could imagine a genius being unrecognised not because she lacked genius but because she could not effectively demonstrate to others that she is a genius. It is difficult for us to say how many individuals exist like this as knowledge of who was in this category would make the category cease to exist.

Therefore, the important question here is whether it is possible for human level intellects to create something that has a greater intelligence to any degree. If it is possible to create a human level AI, then it is possible to create a superspeed conference. If a superspeed conference can design an individually more intelligent AI, call it AI+, then it is only a matter of scale to create a superspeed conference of AI+.[37] That superspeed conference of AI+ can then produce cleverer intelligences than themselves, call them AI++. This process ought to continue until there is a maximal intelligence, $AI+_n$ of which there cannot be a greater intelligence.[38]

There are two ways that this process can fail. We could fail, or be unable, to make the first leap at all; or there is something about the leap between AI+ and AI ++ (or so forth) that does not permit the cycle to continue.[39] Chalmers recognises there could be diminishing returns at each iteration but this does not change the overall structure of ascension to greater intelligences, it only reduces the scale of each leap or lowers the maxima.[40] If we are truly creating a general intelligence like us but cleverer, then it seems arbitrary as to why a later iteration would fail where an earlier one did not. Instead, a future generation of AI+ that failed to be able to improve on itself would be a failure of our design in creation of the first generation of AI+ as it was not truly a superior general intelligence. This is because at some point in the design process for a future intelligence, it failed to have the necessary intelligence

[37] The language of AI, AI+, AI++ is from Chalmers (2010).
[38] Or continue forever if there no maximal intelligence. I am inclined to think that there is because, at least by my account, if it always finds the best answer, it would be maximally intelligent.
[39] There is a third way this could fail which is that improvement is stopped by a barrier than intelligence alone cannot solve. This could be anything from planetary disaster to empirical barriers to progress. This is essentially an argument against the intelligence explosion itself such that I will not dwell on it here. See Chalmers (2010) for a list of, and arguments against, potential defeaters.
[40] Chalmers, "Philosophical Analysis", 26-27.

in some domain (or sufficient generality) to overcome a specific problem. This could occur if we designed something that appeared to be an AI+ but fails in some unforeseen or unrecognised domain that becomes pivotal in design and manufacturing of future AI++. It is for this reason that I seek to analyse this first step, whether we can design an AI+ at all. If we do not get it right, we will not only fail to have a truly superintelligent AI+ but any future $AI+_n$ will also fail to be superintelligent. In the next section, I will analyse whether we can design an AI+ with increased quality of intelligence that also possesses sufficient general intelligence.

5 Can we design a quality superintelligence?

By designing superintelligence, I mean that we write the program ourselves based on how we think it should operate. This kind of programming is called direct programming. The alternative is indirect programming, or machine learning, in which we design through outcomes by providing an algorithm with labelled data. I will return to this kind of programming in the next section.

The aim of a design process is to build a general intelligence algorithm, like one that we possess, but with superior intelligence. General intelligence could have three possible structures. First, it could be a simple and elegant algorithm that scales directly with processing power. If this is true, then the difference in intelligence between a worm and a human is primarily due to the sheer number of neurons. I am not persuaded by such a claim, and AI researchers have found little success in it, although such an idea was championed by some in the early days of artificial

intelligence.[41] More evidence of this is that there are many animals with larger brains than us that are not as intelligent, and humans vary significantly in intelligence but have nearly identically sized brains. Moreover, if true, we ought to expect someone to have designed or discovered such an algorithm by now and run it on whatever available hardware they had. We would be able to quickly know whether the algorithm worked by asking it problems that may take us minutes to solve, but it solves it in days or hours. If it was correct, or at least general enough, we could conclude it was a general intelligence algorithm and then grant it more processing power. Although, it is difficult for me to disprove that such an algorithm could exist, it seems even more difficult to imagine the content of such an algorithm. What could be able to run on a worm, a bird, a tortoise and a human to such differing results? Therefore, I am inclined to believe that this method will not lead us to construction of superintelligence.

The second possibility is that general intelligence is a singular algorithm but it is a very complex one. General intelligence by this account would be an algorithm that controls and instructs a set of domain specific algorithms by allocating parts of a problem to them. Superior general intelligences have greater processing power but they also possess a more sophisticated algorithm that lends itself to better problem solving and other higher cognitive abilities. To improve in general intelligence is improving this algorithm. This seems intuitive and fits well into the description of superintelligences as perfect reasoners that I addressed earlier.

[41] Hubert L Dreyfus, *What Computers Still Can't Do: A Critique of Artificial Reason* (MIT Press, 1992).

However, this is an exceptionally difficult design problem that AI researchers have been wrestling with for decades to little success. It is possible that they lacked a specific set of domain specific algorithms needed for any general intelligence to function but it is unclear why any particular domain should be required. If anything, the current era ought to be the moment when such a system should develop because AI has begun significant inroads into perceptual intelligence like visual and audio recognition among others. If proponents of the Singularity are correct, such a venture ought to result in significant profit such that we ought to have built such an algorithm already.[42]

This is not a reason to believe that such an algorithm cannot be created. It is difficult for me to prove such a claim. The problem for this kind of general intelligence is that we are demanding a great deal of elegance and simplicity where we have little evidence that all problems ought to be so simple. The algorithm must apply extremely generalised rules but be able to achieve specific answers in narrow cases. As it turns out, philosophers have been pondering similar problems in the form of rationality and decision theory. Were we to construct a perfect theory of either, that would be the kind of expression needed for a general intelligence algorithm. It would tell us exactly how much information we need, how to manipulate it and how to handle uncertainty. Unfortunately, philosophers have yet to agree upon answers to the problems in rationality. There are many candidate structures but none have reached consensus or are without flaws. Moreover, none of them are structures actually present in human brains. Some argue that measurements of general intelligence, like IQ, do not correlate with increases in the capacities that we are

[42] Jesse Prinz argues similarly that Chalmers (2010) is too optimistic about the singularity such that it ought to have occurred and therefore we should be sceptical of his arguments. See Prinz (2012).

looking for anyway and that we ought to reject the existence of general intelligence.[43] Our initial reason for believing general intelligence exists was that we had it and could thereby build a superior one. Accepting that general intelligence is something that we do not have would be an implicit rejection of the intelligence explosion cycle that we are attempting to start (or a suggestion of a totally different kind of cycle). Therefore, I am inclined to think that development of this kind of general intelligence is also unlikely.

The final possible structure is that general intelligence is an emergent property of a complicated mess of domain specific intelligences that overlap in particular ways such that they can address any problem. There is no specific general intelligence algorithm that sits atop a hierarchy of intelligences, rather domain specific intelligences overlap one another in a complicated structure that allows general intelligence. I am purposefully vague about exactly what this structure is because, at least for humans, neuroscience has yet to figure it out. As I did in the previous section, I will begin my analysis with human intelligence because we ought to know it better than any hypothetical models of general intelligence that have yet to be built. This allows us to begin with human intelligence and improve upon it, in hopes that we gradually increase individual intelligences until we have increased them all, leading to superintelligence.

5.1 Improving human general intelligence

One way to do this is Human augmentation. This could increase processing power of certain parts of our brain. However, this is something we are already

[43] Susan Greenfield, "The Singularity: Commentary on David Chalmers", *Journal of Consciousness Studies* 19, no. 1-2 (2012): 112-118.

experiencing. Calculators increase our mathematical capacity and smartphones increase our memory capacity. However, neither of these kinds of augmentations appear to modify our intellect in the ways that we might associate with superintelligence. Regardless, any human modification to the speed of parts of our brain will produce very similar results to the whole brain emulation I discussed earlier.

Bostrom discusses this as a pathway to superintelligence but also discusses Whole Brain Emulation, and Chalmers agrees that this is a reasonable possibility.[44] Chalmers believes that there are other ways to develop human level AI through evolutionary methods, that engineers can then improve upon.[45] Bostrom and Shulman are cautiously sceptical of this approach.[46] Regardless, it seems like a better starting point for our analysis.

Therefore, I will begin with the assumption that we develop a human level AI, like Einstein 2.0, and assess what we can do to make it superintelligent. It is important to note that we must do more than improve intelligence in one domain. The most intelligent human is not even a superintelligence in one domain, let alone all of them. However, it seems reasonable enough to assume that we could give Einstein 2.0 a few thousand years of processing time to learn whatever disciplines it is not versed in and therein it may be on the precipice of superintelligence.[47] What

[44] Bostrom and Sandberg, "Roadmap".
[45] Chalmers, "Philosophical Analysis", 7-11.
[46] Nick Bostrom and Carl Shulman, "How Hard is Artificial Intelligence? Evolutionary Arguments and Selection Effects", *Journal of Consciousness Studies*, 19, no. 7-8 (2012): 103-130.
[47] Although this may prove more difficult than I have detailed here. It is possible that individuals that are intelligent in one domain observe the world in ways that fundamentally undermine them in other domains. I will set that concern aside for this paper.

remains is to improve it further into superintelligence. A good starting point for improvement is psychological flaws. Modern psychology has led us to understand that human reasoning and rationality is fundamentally flawed.[48] Almost constantly, we make systematic errors in evaluation of the world that lead to identifiable biases. Let us suppose that we create an Einstein 3.0 which is the improved Einstein 2.0 without some cognitive biases.

The problem with this claim is there are no such peoples without these biases to refer to. We also have no instances of general intelligences that do not have biases like this. Therefore, it is unclear as to whether an intelligence can function at all without biases. These exist in our psychology for a reason. If they are purely detrimental then we ought to expect them to be selected away, or at least to not be present in more intelligent individuals. They are not only present in human psychology but they seem to be perpetuating. This should lead us to doubt the extent to which we can harmlessly remove such cognitive failings. However, there are obviously some cognitive biases that we could be confident in removing. Einstein 3.0 is likely to be better off without racism or logical fallacies. However, it is hard for us to know exactly what part of our psychology leads to such biases or whether they can be removed easily without damaging other parts of our cognition. If detrimental biases are a result in misfires of a system that otherwise uses them well, it may not be possible to remove them without preventing the functioning of the entire system. I am inclined to this interpretation because even the smartest amongst us still succumb to biases so it seems plausible that they are hard to isolate. Bias still plague even the most rigorous scientific systems which indicates how ingrained these

[48] Daniel Kahneman, *Thinking, Fast and Slow.* New York: Farrar, Straus and Giroux, 2011.

failings are. Furthermore, the extent to which we can isolate such biases is contingent on our understanding of what a good general intelligence algorithm looks like. If we lack any particular rubric for what a general intelligence needs to be, or what a good intelligence beyond ours looks like, we will need to try multiple methods and test to see if they are correct. In this next section, I will assess whether we can test if something is a superintelligence or not.

6 Can we test for superintelligence?

If my previous argument is correct then we lack a good understanding of what the internal structure of a superintelligent algorithm ought to look like. This is a common feature of designing anything. At the start of a project, designers will often have the task of making something that fits a purpose rather than fits a specific design. Then they iterate over candidate versions and narrow down potential candidates until they find successful ones. One way to do this is to examine the structure of an algorithm and decide whether it is more or less intelligent. The arguments in the previous section ought to show why this process is difficult. Moreover, even with simple programs, it is often difficult to examine code and conceptualise how it works without running it. This is because computers execute programs so quickly, with so many variables, that it is sometimes near impossible to understand it all.[49] Moreover, it may be very difficult to distinguish between something that is human level intelligence, or slightly more, or slightly less purely by looking at millions of lines of code, in which the difference could be anywhere (or

[49] This is why annotating code is considered as essential for programming as even between humans, code can become unintelligible. This is even worse for programs developed with machine learning in which the final output is written as heuristics rather than any intelligible code.

nowhere). Therefore, the only remaining solution is to differentiate candidate intelligences based on their outputs.

There are many methods of creating candidate intelligences such that it would be impossible for me to properly address each method in this paper. If we cannot produce any of these, through machine learning, or tinkering human level AI or whatever else, then we would have no chance of ever building a superintelligence. Therefore, I will assume that we can produce an arbitrarily large set of candidate superintelligences.[50] If we apply the method of measuring intelligence I detailed in section two, then we could consider something superintelligent if it produces solutions to problems that a superintelligence would also give.

6.1 What is a superintelligent solution?

A flippant answer to the above question is, 'a clever one'. But how can we know what a superintelligent answer is? We are not superintelligences and do not produce superintelligent answers. The answers that we give are ones that correspond to our quality of intelligence. This is the entire reason why we want to build a superintelligence in the first place. We want solutions that we have not thought of. Moreover, we want it to solve the future design problems of making higher level intelligences that we cannot solve and have not yet encountered. So, for something to be a superintelligence, it should produce answers that the best thinkers are not producing now. Quality superintelligences ought to be capable of this, in the same way that we expect humans with higher quality intelligences to produce answers that others do not. If the only answers that a superintelligence is generating are answers

[50] This is not necessarily all at once, candidates in this set could be a result of tinkering parts of an AI one change at a time over a large period of time.

that human intelligences also give, then it would be good evidence that it is not a superintelligence.

It is important to establish how a superintelligent answer appears from the outside. If we think that we have the correct answer, then we could be making two claims. First, that of all the options we could come up with, this was the best. Second, any options that we did not think of were not good answers. That means that a superintelligent solution could come in one of three forms. It could be something that we had ruled out as incorrect; something that we had not thought of at all; or exactly the answer that we had. Many of our candidates will not be superintelligences, with most having lower than human level intelligence in some ways. The problem is telling the difference between an answer that is cleverer than ours but we thought it was wrong or did not address it, or an answer that is worse than ours but we thought was wrong or did not address it. Both appear the same from the outside. Without some kind of extra measure of intelligence, we are in a position of ignorance because we cannot tell between something that is stupid or clever. This is a common position of ignorance. For example, if I have absolutely no knowledge about cars and I were asked about the speed of a car relative to the average I would be equally likely to think it was faster as I would that it is slower (*ceteris paribus*). If I were then given the speed of the car, the truth would be indistinguishable from a lie. By this reasoning, a superintelligent answer should be *prima facie* indistinguishable from a less intelligent answer. This renders us unable to tell if any candidate has superintelligence, human intelligence, or a less than human intelligence. In the next sections, I will address possible methods of performing this differentiation.

6.2 The 'I know it when I see it' method

One way of knowing that something is intelligent is by simply recognising it. I might say that, although I may not be able to tell you why, I can tell something is intelligent because '*I know it when I see it*'. The method stipulates that even though I may not be able to necessarily produce an answer that is above my level of intelligence, I can recognise when something is more intelligent than what I could produce. This seems reasonably intuitive. We are often able to understand and point out when someone is saying or doing something intelligent. Consider being lost in a problem where you cannot figure out a solution until a clever friend gives you a brilliant answer that you immediately recognise to be a good idea. You could not have thought of it yourself, or at least you did not think of it, but you know immediately that it is a clever idea. It seems that you can recognise that a solution is intelligent without having previously generated that solution. Similarly, if we are presented with a problem we understand quite well, it appears we can determine when an answer is wrong even when we have yet to find the right answer. Both provide a potential answer to the problem. If we can recognise when an answer is intelligent or unintelligent, then we are able to identify which candidates are superintelligences.

We need to be clear about exactly what is going on in these cases. One possible explanation is that we are verifying the answer by checking that it is correct. There are many ways to do this and I will return to verification later. The method here is that we can understand that something is a good answer by glancing at it rather than by testing in some way that it works. One way we might do this is by inferring intelligence through something about the structure rather than an explicit

verification of its content. This seems intuitive enough. We often claim that something appears intelligent and use this as a heuristic to claim that it is intelligent. The problem is that it is unclear what structural rules a superintelligent answer ought to follow. Were it to have an obvious structure, surely that is something we would be thinking about already and making our own answers follow that structure. Structure can also be deeply misleading. It can allow us to only consider answers that are structurally like our own, or ones we already consider as intelligent.

For this to be a good measure of superintelligence, the structure we associate with intelligence must actually be effective at recognising it. However, our recognition of intelligence in others is often sorely mistaken. [51] Especially with ideas that are different to our own. Humanity has a long history of opposition, even amongst experts, to correct but controversial ideas. We have a proven ability to see the correct answer and vehemently believe it is false and worthy of ridicule. It is only upon repeated clashes with the evidence from the real world that controversial ideas begin to gain traction. This should lead us to seriously doubt our capacity to perform this identification role, especially in instances in which the intelligence could be radically dissimilar to us. A good example of this is the phenomenon of 'computer-moves' in Chess. These are moves performed by high level Chess computers that to human observers look like bizarre, and oftentimes bad, moves but are in fact good moves. The only reason high level Chess players recognise that they are good moves is because they know that the Chess computer is better than them, not because they can grasp why the move is strategically useful. If we hijacked the system at that moment, unbeknownst to the Chess players, and actually played a bad move. Those

[51] Some common indicators that we often use are particularly bad, like accent or vocabulary.

same players would claim that the move was a good one and equally have no understanding why. The same phenomenon happens between human players as well. Were I to play against Garry Kasporov, I would not understand his moves (and they could appear very strange) but I would assume that they are good ones purely because a better Chess player is making them. Indeed, if we want our superintelligences to be powerful and impactful we ought to expect them to give strange answers to questions because we are likely already considering the obvious ones. Therefore, we ought to think that recognition will not be a good method but some kind of verification in the real world or otherwise is a far better measure of intelligence.

6.3 Verification

A good starting point for verification are instances in which we can understand the logic of the answer as it is presented and work out that the solution is a correct one. The person giving us the answer either implicitly, through their answer, or explicitly tells us what their reasoning is. There are a few problems with this kind of verification. First, it is unclear as to why we should assume that the logic a superintelligence uses to produce an answer is obvious or even communicable to humans. It seems plausible to me that the best option in a given situation may defy elegance, simplicity, be deeply counterintuitive or difficult to explain. The latter seems especially pertinent in cases where we develop potential superintelligences but they fail to advance in communicative ability, or any other specific domain required to persuade us of their correctness. This happens in the real world too. Often the solutions to the world's problems, devised by those with the best knowledge and understanding, fail because they are so counterintuitive or difficult to

grasp that the public (or even their peers) consider them as bad options. The problem here is that we will be forced to engage with candidate intelligences and differentiate between; good ones that produce answers with complicated reasoning, and bad ones that produce answers with complicated reasoning. This puts us in a similar position of ignorance that we began with.

Second, although there are certainly some problems in which verification is easy, there are many problems in which verification is extremely difficult. The easy problems are those in which computers are already excelling. Consider mathematical problems. A computer was able to find a counterexample to one of Euler's conjectures.[52] The counterexample is extremely easy to verify, one can even do it without a calculator. All one needs to do is to add up the sum and verify that it is in the correct structure. However, it seems to me that maths problems are the exception rather than the norm. For problems that are about the real world, the only verification is the real world or a complicated simulation. For some problems, a simple simulation is good enough. When given the great idea from a friend, I might be able to verify it purely because the terms of the problem are very simple. I imagine the relevant parts of the world and play out their solution. If it works as they say, I conclude that the idea is a good one. Failure in the real world occurs when the rules of the simulation do not actually apply to the real world because they are too simplistic or the wrong rules. This is a common mistake. It is so common that a mark of higher intelligence is an understanding of precisely this fact and responding by sketching the world in less simplistic terms. If our verification process is only appealing to our simplistic understanding of the rules of the world, then we are likely

[52] L. J. Lander and T. R. Parkin, "A Counterexample to Euler's Sum of Powers Conjecture", *Mathematics of Computation* 21, no. 97 (1967): 101-101.

to be mistaken about a great many problems. The very problems that humans verify easily are precisely those that we do not need a superintelligence to solve. However, this argument somewhat applies between individuals as well, yet we seem to be capable of verifying each other. The reason for this is that our verification is assisted by exposing our actions and reasoning to the real world, to affirm whether we are correctly assessing reality or not.

6.4 Exposure to the real world

Our normal intelligence verifying process is that we allow individuals to progress through the world and demonstrate their intelligence in naturally ascending instances of impact. As a child, we can demonstrate benign levels of intellect by bettering our peers. This then allows us to have access to more and more impactful demonstrations of intellect. The process limits the power to cause harm to those that have demonstrated the greatest ability to manage those harms.[53]

This strategy is problematic for AI for three reasons. First, there will likely be an extremely high number of candidates such that testing in the real world is extremely cumbersome. It is difficult to allow exposure to the real world for billions of humans to test their abilities, let alone potentially equally many artificial intelligences.[54] Second, it is difficult to create controlled testing situations because we are trying to find superintelligence rather than intelligence. Most problems that we are presented with are likely optimised to the highest, or close to highest, degree by a normal human. It is not as if only geniuses can drive to work, or engage socially, or perform

[53] At least when it is working well.
[54] One of the many tragedies of a world filled with poverty is that we have almost certainly left countless geniuses to toil in fields or tragically die young.

most everyday problem-solving tasks. In the same way that you would not test whether someone is a maths genius with basic addition problems, you need difficult problems to challenge candidates such that you can differentiate whether they are human level intelligences or superintelligences. This alludes to the third problem, it is extremely dangerous.

Any of these intelligences could have the capacity to cause extreme amounts of harm, whether it be due to a radical superintelligence or by misalignment such that it performs more stupid courses of action than a human. We also ought to expect many of the latter kinds of systems because many more of the intelligences will fail to be superintelligences than those that succeed.[55] We could probably choose not to test the radically stupid ideas from candidates but we risk excluding superintelligences altogether. As I demonstrated earlier, the answers from a superintelligence will likely appear to be less than optimal or at least strange. If they are not actually correct, then these options could be extremely dangerous. Our current systems often fail to prevent humans causing harm through mistakes. It would be at least as difficult, if not more so, to prevent colossal harm from rogue candidates. Moreover, if Bostrom is right about superintelligences, letting them have full contact with the real world could lead superintelligences to hide themselves and escape our control. That said, I do not doubt that letting candidates run wild in the world would identify which are superintelligent.[56] However, such an action could be impermissibly irresponsible

[55] If not, then superintelligence is a much easier problem that anyone thought.
[56] There are serious concerns as to with whether we could set goals for these intelligences such that they would not incidentally cause serious ill in the world. If we cannot find an answer to the problem, the control problem, then it may not even be possible to measure them on whether they do good in the world. Rather, only on goal completion. See Bostrom (2012).

because it could cause significant damage to humanity. One way we might avoid this problem is by testing in simulated worlds rather than the real world.

6.5 Simulated worlds

This method has been used for many AIs but has had recent success for the board game Go, with Deepmind's AlphaGo.[57] As with superintelligences, we are ignorant of what the best move should be but we have an accurate depiction of what the end state should be, a victory by the ruleset of the game. This allowed engineers to run millions of games where the AI played itself and tested various heuristics, methods and strategies to slowly perfect the algorithm until it could defeat any human opponent. This method overcomes our ignorance of what superintelligent moves are because we have certainty of the end state. Therefore, simulating the world and allowing potential intelligences to run wild in those worlds would allow us to see if they are superintelligent. It could display its intelligence by solving the simulated world's problems or by some other indicator like enslaving the human race.[58] Either way, it seems plausible enough that we could gauge whether a candidate was superintelligent by how it interacted with these worlds. However, there are two reasons to doubt our ability to create such a simulation.

The first is simply computing power. The ruleset of Chess is known and it is computationally trivial to simulate that ruleset. However, the real world is computationally very difficult to simulate to any meaningful level of detail.

[57] David Silver et al., "Mastering the Game of Go with Deep Neural Networks and Tree Search", *Nature* 529, no. 7587 (2016): 484-489.
[58] This does provide the additional hazard of a superintelligence realising it is in a simulated world and tricking us into thinking that it is safe to use in the real world when it is not. See Armstrong, Sandberg and Bostrom (2012).

Simulating a brain is difficult enough because of the magnitude of interactions and physical complexity. The world not only contains many brains but also many other systems with as much complexity. To test candidate intelligences, we would need to run many simulations, for many intelligences, for potentially large amounts of time. Although I have already assumed significant computational power to simulate brains, the level of computational power for simulations is considerably greater than that.[59]

Second, it is very difficult to separate what needs to be simulated from what does not. The strength of the Go simulation was perfect ruleset emulation such that it allowed engineers to explore many potential strategies, knowing that they would execute perfectly in the real world. If we choose to only simulate certain parts of the real world, on pain of computational impossibility, we lose this certainty. We would be training our intelligences purely with respect to the rules of the simulated world rather than the rules of the real world. Moreover, this also hinders us from gauging which candidates are general intelligences because some intelligences succeed in simulated worlds because they have domain specific intellects that are well suited to them, but not to the real world. Therefore, unless we can produce such simulations accurately we will be unable to simulate them at all or run the risk of approving extremely dangerous intelligences for action in the real world.

Nick Bostrom has discussed a similar approach to this, instead suggesting we can plausibly evolve potential intelligences within simulated spaces.[60] He also concluded that computing power would be a significant limitation on simulated

[59] Bostrom discusses computational requirements to simulate the world as compared to individual brains in the context of whether we are simulated. He concludes that a totally detailed world is considerable orders of magnitude more difficult to simulate than a single brain. See Bostrom (2003).
[60] Bostrom and Shulman, "Evolutionary Arguments", 103-130.

worlds. He, as well as Chalmers in another paper, argue that engineers may be able to properly simulate intelligence favouring conditions within a simulated world to promote the evolution of intelligent life forms.[61] In principle, a method like this ought to avoid some of these selection problems. If our intelligence favouring conditions are precisely correct then we could know for sure that an end result was superintelligent. However, it is not obvious what the required set of conditions are. They ought to be such that the survival of an organism depends on solving problems in an intelligent and later superintelligent way. If the conditions can be sufficiently overcome with human level intelligence, then the selection mechanism will only progress to human level of intelligence. This is precisely why most organisms on earth do not possess human level intelligence because they do not need it to solve the problems which are presented to them. To overcome this, selection pressure must be such that superintelligent solutions have a strong evolutionary benefit that intelligent solutions do not. The limitation once again is that we do not know what a superintelligent solution looks like or even which problems a superintelligence could give us a better solution. Therefore, we could not be sure that we were placing the right kinds of intelligence pressure on the system to guarantee superintelligence. If we cannot be sure that an output is superintelligent, then we will need to test that system and my previous objections apply. In this next section, I will address one final possible method for verification, in which we determine something is a superintelligence by verifying it using tests which we know a superintelligence would succeed at.

[61] Although Chalmers argues that this will allow us to develop the first human level AI, not necessarily a superintelligence. He then claims that we can improve upon such an intelligence through more evolution or through incremental design improvement. See Chalmers (2010).

6.6 Cross-verification

Suppose that we had a million candidate general intelligences and we wanted to test whether they were superintelligences. We also take my above arguments seriously and conclude that we are unlikely to correctly assess answers from a superintelligence to questions that we do not know the perfect solution to, like "what is the best way to reduce crime?". However, we do know that there are some tasks in which a true general intelligence ought to excel at, even if it has never encountered the problem. Therefore, we decide to tell each candidate intelligence the rules of Chess and have it play a few games with a rotation of Chess Grandmasters. If a candidate can beat them in Chess, then we ought to have good evidence that it is a general superintelligence. We could also perform this same kind of function by asking the intelligences to find proofs for problems in maths and verifying the proofs they produce.

The problem with this method is that we are tacitly assuming that the intelligences are general intelligences. Consider the space of intelligences that could be good at Chess. These include, Chess intelligences, Chess and Checkers intelligences, board game intelligences and a host of other potential intelligences. It seems no more likely that we develop a general intelligence rather than any of those. Moreover, within the space of candidate intelligences, there are a host of deeply flawed candidates that are good at Chess, or Chess-like, games but systematically fail in other aspects. Therefore, we would be unjustified to conclude that something was a general superintelligence rather than an intelligence that is good at Chess, simply because it is good at Chess. Additionally, Superintelligences may also be at a

disadvantage to their less general counterparts as they may require more resources to compute because they could be using less focused algorithms.

This method becomes more promising if we add some arbitrarily large set of verification tasks (assuming we have enough of these kinds of problems). Suppose there is some number of superintelligences hidden amongst a million candidate superintelligences. If there is a finite probability of any given intelligence being a superintelligence, there ought to also be a finite probability that any intelligence is not a general intelligence but passes all the cross-verification problems (like instant proficiency in Chess, mathematics and Go). The more tasks we add, the probability of any of these intelligences in the pool decreases whilst the chance of a general intelligence does not change. Therefore, we would be able to place a high likelihood on an intelligence that passes verification to also be a superintelligence.

If the above measure of probabilities is correct then we ought to have good evidence that anything that passes the tests is likely to be a general intelligence, rather than something that is only clever in the testing domains. However, I am not confident that the probability must be like this. It seems very difficult to say what the probability of a superintelligence ought to be within our set. Depending on what general intelligence is, it is also plausible that the probability of a general superintelligence is always less than the probability of something that has a specific number of domain intelligences.

I am unsure what we are to conclude about these probabilities. Future AI researchers may find themselves in the same conundrum. Until we discover which

method actually increases intelligence uniformly, rather than decreasing it, it is hard to say what the probability of those systems arising ought to be. Therefore, it would be unwise for us to conclude something was a superintelligence based on a probabilistic argument in which the probabilities are unknown.

7 What should we expect from superintelligence?

If the arguments of the previous two sections have been successful then we ought to conclude that the development of superintelligence is very difficult. This is because we do not know how to identify superintelligence by its design, or by its outputs in controlled settings. These obstacles are extremely hard to overcome because our knowledge of superintelligence is limited to our intelligence. My arguments ought to also apply to any future AI_n in development of later AI_{n+1} in an intelligence explosion situation. This is because they relate to the construction of greater intelligences, not only the creation of an intelligence that is greater than humans. Therefore, I have shown that it is, at least, extremely difficult to build something with greater qualitative intelligence than itself. However, some of this is to be expected. I suspect very few will claim that construction of a superintelligence ought to be easy.

Even if it is difficult, it is hard for me to claim that we cannot build a superintelligence altogether. In some sense, I already refuted that claim by admitting the possibility of speed superintelligence in the form of superspeed conferences. So, what is unique to quality superintelligences? I argued that the intelligence explosion

could only be facilitated by quality superintelligences that were also general intelligences. Even if this is correct, one could still claim than an intelligence explosion will occur but of a different kind. A superspeed conference is likely to assist in development and discovery of better computer hardware, which allows us to build a bigger conference and so forth. It also allows us to free up more labour to do other work like gathering resources or assembling computers and progress humanity that way. One could even argue this is already happening. Humanity has already grown in intelligence by multiplying and educating more people. As the number of people doing intelligent labour grew, we developed solutions to increase our ability to support more humans and so forth. If we are in an intelligence explosion already, we do not seem to be acquiring some of the characteristics of superintelligences I discussed in this paper.

This is because, as I demonstrated, speed superintelligences are different to quality intelligences. Bostrom, Kurzweil and Chalmers either explicitly argue this or seem to implicitly recognise this. However, they tend to argue that speed superintelligences will lead to quality superintelligences, or lead to a similar character. However, speed superintelligences comprise of individual human agents that bear the weight of human values and failings. So, we ought not to expect them to exhibit the hyper-rationality and pseudo-omniscience that seems to come with quality superintelligences. At the very least, they ought to be very different. This leads to less dramatic advancement in intelligence but can also be useful. Speed superintelligences are less likely to accidentally destroy humanity in response to a request for too many paperclips.

However, quality superintelligences often have the kind of insurmountable intellectual power that might lead them to accidentally destroy us. Why are we attracted to these expectations? I want to suggest that the definition of superintelligence is more malleable. Becoming more intelligent is not necessarily an inexorable rise to a nightmarish Bayesian god, we have only defined it as such.[62] Intelligence is so vague that it seems difficult to argue that there is an objective standard for what superintelligence must mean.[63] I have been admittedly vague in my discussion about how intelligent reasoning works. This is because there are many ways to do it and it is yet unanswered which is correct, or which one we ought to use.[64] Furthermore, we portray human intelligence as the source of all technological and scientific progress, where ape intelligence failed to do any of this. From there it seems logical to imagine that if we extend that difference, we achieve as much of a contrast. The problem is that we give too much credit to human intelligence. A significant portion of human progress was likely sheer luck, serendipity or brute force from many humans blindly testing many things. Moreover, we tend to anthropomorphise measurements of intelligence that make it seem as if humans are vastly more intelligent than our peers. Maybe animals think us fools for failing in all the domains where they succeed. Therefore, our intelligence, especially general intelligence, may not be as powerful as we may think.

Therein, the problem for superintelligence as a concept is that it is an extension of intelligence. To the extent we chose the burden for intelligence, we also chose it

[62] Bostrom, *Superintelligence*, 12-13.
[63] Marcus Hutter discusses a few of the difficulties with intelligence as they apply to superintelligence in a response to Chalmers (2010). See Hutter (2012).
[64] Shane Legg and Marcus Hutter, "Universal Intelligence: A Definition of Machine Intelligence", *Minds and Machines* 17, no. 4 (2007): 391-405.

for superintelligence. The arguments in this paper do not show that we cannot build superintelligences, only that we cannot build *some* superintelligences. I suggest that this is a result of superintelligence being ill-defined. If true, the singularity is a bit less exciting. It does not march us towards an omniscient superintelligence, rather to very complicated but potentially powerful intelligences. Instead of approaching superintelligences as automatically hyper-rational agents, the character of eventual superintelligences could be far more humanlike, or at least more fathomable. The malleability of the meaning of superintelligence grants us more scope to shape how it may eventually act. We ought to place more focus on varied conceptions of intelligence that may be cleverer in some ways and worse in others. A drift away from quality general intelligences allows us to better explore this space. Although we may not be able to start a quality intelligence explosion, we can probably build a superspeed conference of some sort. However, if we were to fill it with psychopathic geniuses, I would still be worried.

Bibliography

Allis, Victor. *Searching for Solutions in Games and Artificial Intelligence*. Maastricht: Rijksuniversiteit Limburg, 1994.

Armstrong, Stuart, Nick Bostrom and Anders Sandberg. "Thinking Inside the Box: Controlling and Using an Oracle AI". *Minds and Machines* 22, no. 4 (2012): 299-324.

––– and Nick Bostrom, Carl Shulman. “Racing to the Precipice: A Model of Artificial Intelligence Development”. *Technical Report #2013-1,* Future of Humanity Institute, Oxford University (2013): 1-8.

Block, Ned. "Psychologism and Behaviorism". *The Philosophical Review* 90, no. 1 (1981): 5.

Bostrom, Nick. "How Long Before Superintelligence?". *International Journal of Future Studies* 2 (1998): 1.

––– "Are We Living in a Computer Simulation?". *The Philosophical Quarterly* 53, no. 211 (2003): 243-255.

––– and Anders Sandberg. "Converging Cognitive Enhancements". *Annals of the New York Academy of Sciences* 1093, no. 1 (2006): 201-227.

––– and Anders Sandberg. “Whole Brain Emulation: A Roadmap”. *Technical Report #2008-3*, Future of Humanity Institute, Oxford University (2008).

––– and Carl Shulman. “How Hard is Artificial Intelligence? Evolutionary Arguments and Selection Effects”. *Journal of Consciousness Studies*, 19, no. 7-8 (2012): 103-130.

––– "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents". *Minds and Machines* 22, no. 2 (2012): 71-85.

––– *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, 2013.

Chalmers, David. "The Singularity a Philosophical Analysis". *Journal of Consciousness Studies* 17, no. 9-10 (2010): 7-65.

––– "The Singularity: A Reply to Commentators". *Journal of Consciousness Studies* 19, no. 7-8 (2012): 141-167.

Chomsky, Noam. "Turing on the "Imitation Game"". In *Parsing the Turing Test*. Dordrecht: Springer, 2009.

Dreyfus, Hubert L. *What Computers Still Can't Do: A Critique of Artificial Reason*. MIT Press, 1992.

Good, Irving John. "Speculations Concerning the First Ultraintelligent Machine". *Advances in Computers* 6 (1966): 31-88.

Greenfield, Susan. "The Singularity: Commentary on David Chalmers". Journal of Consciousness Studies 19, no. 1-2 (2012): 112-118.

Harnad, Steven. "Minds, Machines and Turing: The Indistinguishability of Indistinguishables". *Journal of Logic, Language, and Information* 9, no. 4 (2000): 9.

Hutter, Marcus. "Can Intelligence Explode?". *Journal of Consciousness Studies* 19, no. 1-2 (2012): 143-166.

Kahneman, Daniel. *Thinking, Fast and Slow.* New York: Farrar, Straus and Giroux, 2011.

Kripke, Saul A. *Wittgenstein on Rules and Private Language*. Cambridge, Mass: Harvard University Press, 1982.

Kurzweil, Ray. *The Singularity is Near: When Humans Transcend Biology*. New York: Viking, 2005.

Lander, L. J., and T. R. Parkin. "A Counterexample to Euler's Sum of Powers Conjecture". *Mathematics of Computation* 21, no. 97 (1967): 101-101.

Legg, Shane, and Marcus Hutter. "Universal Intelligence: A Definition of Machine Intelligence". *Minds and Machines* 17, no. 4 (2007): 391-444.

Müller, Vincent C. "Risks of General Artificial Intelligence". *Journal of Experimental & Theoretical Artificial Intelligence* 26, no. 3 (2014): 297-301.

Prinz, Jesse. "Singularity and Inevitable Doom". *Journal of Consciousness Studies* 19, no. 7-8 (2012): 7-8.

Searle, John R. "Minds, Brains, and Programs". *The Behavioral and Brain Sciences* 3 (1980): 417-457.

Shannon, Claude E. "Programming a Computer for Playing Chess". *Philosophical Magazine* 41, no. 304 (1950).

Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, and Julian Schrittwieser et al. "Mastering the Game of Go with Deep Neural Networks and Tree Search". *Nature* 529, no. 7587 (2016): 484-489.

Turing, A. M. "Computing Machinery and Intelligence". *Mind* no. 236 (1950): 433-460.