Modeling and Interpreting Expert Disagreement About Artificial Superintelligence

Seth D. Baum, Anthony M. Barrett, and Roman V. Yampolskiy
Global Catastrophic Risk Institute, PO Box 40364, Washington, DC 20016, USA
http://gcrinstitute.org,
E-mail: seth@gcrinstitute.org

Keywords: artificial superintelligence, expert judgment, risk analysis

Received: Avgust 31, 2017

Artificial superintelligence (ASI) is artificial intelligence (AI) with capabilities that are significantly greater than human capabilities across a wide range of domains. A hallmark of the ASI issue is disagreement among experts. This paper demonstrates and discusses methodological options for modeling and interpreting expert disagreement about the risk of ASI catastrophe. Using a new model called ASI-PATH, the paper models a well-documented recent disagreement between Nick Bostrom and Ben Goertzel, two distinguished ASI experts. Three points of disagreement are considered: (1) the potential for humans to evaluate the values held by an AI, (2) the potential for humans to create an AI with values that humans would consider desirable, and (3) the potential for an AI to create for itself values that humans would consider desirable. An initial quantitative analysis shows that accounting for variation in expert judgment can have a large effect on estimates of the risk of ASI catastrophe. The risk estimates can in turn inform ASI risk management strategies, which the paper demonstrates via an analysis of the strategy of AI confinement. The paper find the optimal strength of AI confinement to depend on the balance of risk parameters (1) and (2).

Povzetek: Predstavljena je metoda za modeliranje in interpretiranje razlik v mnenjih ekspertov o superinteligenci.

1 Introduction

Artificial superintelligence (ASI) is artificial intelligence (AI) with capabilities that are significantly greater than human capabilities across a wide range of domains. If developed, ASI could have impacts that are highly beneficial or catastrophically harmful, depending on its design

A hallmark of the ASI issue is disagreement among experts. Experts disagree on if ASI will be built, when it would be built, what designs it would use, and what its likely impacts would be.[1] The extent of expert disagreement speaks to the opacity of the underlying ASI issue and the general difficulty of forecasting future technologies. This stands in contrast with other major global issues, such as climate change, for which there is extensive expert agreement on the basic parameters of the issue (Oreskes 2004). Expert consensus does not guarantee that the issue will be addressed—the ongoing struggle to address climate change attests to this—but it does offer direction for decision making.

In the absence of expert agreement, those seeking to gain an understanding of the issue must decide what to believe given the existence of the disagreement. In some cases, it may be possible to look at the nature of the disagreement and pick sides; this occurs if other sides clearly have flawed arguments that are not worth giving any credence to. However, in many cases, multiple sides of a disagreement make plausible arguments; in these cases, the thoughtful observer may wish to form a belief that in some way considers the divergent expert opinions.

This paper demonstrates and discusses methodological options for modeling and interpreting expert disagreement about the risk of ASI catastrophe. The paper accomplishes this by using a new ASI risk model called ASI-PATH (Barrett and Baum 2017a; 2017b). Expert disagreement can be modeled as differing estimates of parameters in the risk model. Given a set of differing expert parameter estimates, aggregate risk estimates can be made using weighting functions. Modeling expert disagreement within the context of a risk model is a method that has been used widely across a range of other contexts; to our knowledge this paper marks the first application of this method to ASI.

The paper uses a well-documented recent disagreement between Nick Bostrom and Ben Goertzel as an illustrative example—an example that is also worthy of study in its own right. Bostrom and Goertzel are both longstanding thought leaders about ASI, with lengthy research track records and a shared concern with the societal impacts of ASI. However, in recent publications, Goertzel (2015; 2016) expresses significant

[1] On expert opinion of ASI, see Baum et al. (2011), Armstrong and Sotala (2012), Armstrong et al. (2014), and Müller and Bostrom (2014).

disagreement with core arguments made by Bostrom (2014). The Bostrom-Goertzel disagreement is notable because both of them are experts whose arguments about ASI can be expected to merit significant credence from the perspective of an outside observer. Therefore, their disagreement offers a simple but important case study for demonstrating the methodology of modeling and interpreting expert disagreement about ASI.

The paper begins by summarizing the terms of the Bostrom-Goertzel disagreement. The paper then introduces the ASI-PATH model and shows how the Bostrom-Goertzel disagreement can be expressed in terms of ASI-PATH model parameters. The paper then presents model parameter estimates based on the Bostrom-Goertzel disagreement. The parameter estimates are not rigorously justified and instead are intended mainly for illustration and discussion purposes. Finally, the paper applies the risk modeling to a practical problem, that of AI confinement.

2 The Bostrom-Goertzel disagreement

Goertzel (2015; 2016) presents several disagreements with Bostrom (2014). This section focuses on three disagreements of direct relevance to ASI risk.

2.1 Human evaluation of AI values

One disagreement is on the potential for humans to evaluate the values that an AI has. Humans would want to diagnose an AI's values to ensure that they are something that humans consider desirable (henceforth "human-desirable"). If humans find an AI to have human-undesirable values, they can reprogram the AI or shut it down. As an AI gains in intelligence and power, it will become more capable of realizing its values, thus making it more important that its values are human-desirable. A core point of disagreement concerns the prospects for evaluating the values of AI that have significant but still subhuman intelligence levels. Bostrom indicates relatively low prospects for success at this evaluation, whereas Goertzel indicates relatively high prospects for success.

Bostrom (2014, p.116-119) posits that once an AI reaches a certain point of intelligence, it might adopt an adversarial approach. Bostrom dubs this point the "treacherous turn":

> *The treacherous turn*: While weak, an AI behaves cooperatively (increasingly so, as it gets smarter). When the AI gets sufficiently strong–without warning or provocation–it strikes, forms a singleton [i.e., takes over the world], and begins directly to optimize the world according to the criteria implied by its final values. (Bostrom 2014, p.119)

Such an AI would not have durable values in the sense that it would go from acting in human-desirable ways to acting in human-undesirable ways. A key detail of the treacherous turn theory is that the AI has values that are similar to, but ultimately different from, human-desirable values. As the AI gains intelligence, it goes through a series of stages:

1. At low levels of intelligence, the AI acts in ways that humans consider desirable. At this stage, the differences between the AI's values and human values are not important because the AI can only complete simple tasks that are human-desirable.
2. At an intermediate level of intelligence, the AI realizes that its values differ from human-desirable values *and* that it if it tried deviating from human-desirable values, humans would reprogram the AI or shut it down. Furthermore, the AI discovers that it can successfully pretend to have human-desirable values until it is more intelligent.
3. At a high level of intelligence, the AI takes control of the world from humanity so that humans cannot reprogram it or shut it down, and then pursues its actual, human-undesirable values.

Goertzel provides a contrasting view, focusing on Step 2. He posits that an AI of intermediate intelligence is unlikely to successfully pretend to have human-desirable values because this would be too difficult for such an AI. Noting that "maintaining a web of lies rapidly gets very complicated" (Goertzel 2016, p.55), Goertzel posits that humans, being smarter and in control, would be able to see through a sub-human-level AI's "web of lies". Key to Goertzel's reasoning is the claim that an AI is likely to exhibit human-undesirable behavior *before* it (A) learns that such behavior is human-undesirable and (B) learns how to fake human-desirable behavior. Thus, Step 2 is unlikely to occur—instead, it is more likely that an AI would either have actual human-desirable values or be recognized by humans as faulty and then be reprogrammed or shut down.

Goertzel does not name his view, so we will call it the sordid stumble:

> *The sordid stumble*: An AI that lacks human-desirable values will behave in a way that reveals its human-undesirable values to humans before it gains the capability to deceive humans into believing that it has human-desirable values.

It should be noted that the distinction between the treacherous turn and the sordid stumble is about the AI itself, which is only one part of the human evaluation of the AI's values. The other part is the human effort at evaluation. An AI that is unskilled at deceiving humans could still succeed if humans are not trying hard to notice the deception, while a skilled AI could fail if humans are trying hard. Thus, this particular Bostrom-Goertzel debate covers only one part of the AI risk. However, it is still the case that, given a certain amount of human effort at evaluating an AI's values, Bostrom's treacherous turn suggests a lower chance of successful evaluation than Goertzel's sordid stumble.

2.2 Human creation of human-desirable AI values

A second disagreement concerns how difficult it would be for humans to give an AI human-desirable values. If an AI's values are human-desirable, then it is not crucial whether humans can evaluate them, because humans would not want to reprogram the AI or shut it down. As the AI gains in intelligence and power, it would simply take more and more human-desirable actions. Bostrom indicates relatively low prospects for success for humans to give AIs human-desirable values, whereas Goertzel indicates relatively high prospects for success.

Bostrom (2014) argues that AIs are likely to have human-undesirable final goals because these goals are more complex:

> There is nothing paradoxical about an AI whose sole final goal is to count the grains of sand on Borcay, or to calculate the decimal expansion of pi, or to maximize the total number of paperclips that will exist in its future light cone. In fact, it would be *easier* to create an AI with simple goals like these than to build one that had a human-like set of values and dispositions (Bostrom 2014, p.107).

The logic of the above passage is that creating an AI with human-desirable values is more difficult and thus less likely to occur. Goertzel (2016), citing Sotala (2015), refers to this as the difficulty thesis:

> *The difficulty thesis*: Getting AIs to care about human values in the right way is really difficult, so even if we take strong precautions and explicitly try to engineer sophisticated beneficial goals, we may still fail (Goertzel 2016, p.60).

Goertzel (2016) discusses a Sotala (2015) argument against the difficulty thesis, which is that while human values are indeed complex and difficult to learn, AIs are increasingly capable of learning complex things.Per this reasoning, giving an AI human-desirable values is still more difficult than, say, programming it to calculate digits of pi, but it may nonetheless be a fairly straightforward task for common AI algorithms. Thus, while it would not be easy for humans to create an AI with human-desirable values, it would not be extraordinarily difficult either. Goertzel (2016), again citing Sotala (2015), refers to this as the weak difficulty thesis:

> *The weak difficulty thesis*. It is harder to correctly learn and internalize human values, than it is to learn most other concepts. This might cause otherwise intelligent AI systems to act in ways that went against our values, if those AI systems had internalized a different set of values than the ones we wanted them to internalize.

A more important consideration than the *absolute* difficulty of giving an AI human-desirable values is its *relative* difficulty compared to the difficulty of creating an AI that could take over the world. A larger relative ease of creating an AI with human-desirable values implies a higher probability that AI catastrophe will be avoided for any given level of effort put to avoiding it.

There is reason to believe that the easier task is giving an AI human-desirable values. For comparison, every (or almost every) human being holds human-desirable values. Granted, some humans have more refined values than others, and some engage in violence or other antisocial conduct, but it is rare for someone to have pathological values like an incessant desire to calculate digits of pi. In contrast, none (or almost none) of us is capable of taking over the world. Characters like Alexander the Great and Genghis Khan are the exception, not the rule, and even they could have been assassinated by a single suicidal bodyguard. By the same reasoning, it may be easier for an AI to gain human-desirable values than it is for an AI to take over the world. This reasoning does not necessarily hold, since AI cognition can differ substantially from human cognition, but it nonetheless suggests that giving an AI human-desirable values may be the easier task.

2.3 AI creation of human-desirable AI values

A third point of discussion concerns the potential for an AI to end up with human-desirable values even though its human creators did not give it such values. If AIs tend to end up with human-desirable values, this reduces the pressure on the human creators of AI to get the AI's values right. It also increases the overall prospects for a positive AI outcome. To generalize, Bostrom proposes that AIs will tend to maintain stable values, whereas Goertzel proposes that AIs may tend to evolve values that could be more human-desirable.

Bostrom's (2014) thinking on the matter centers on a concept he calls goal-content integrity:

> *Goal-content integrity*: If an agent retains its present goals into the future, then its present goals will be more likely to be achieved by its future self. This gives the agent a present instrumental reason to prevent alteration of its final goals (Bostrom 2014, p.109-110).

The idea here is that an AI would seek to keep its values intact as one means of realizing its values. At any given moment, an AI has a certain set of values and seeks to act so as to realize these values. One factor it may consider is the extent to which its future self would also seek to realize these values. Bostrom's argument is that an AI is likely to expect that its future self would realize its present values more if the future self retains the present self's values, regardless of whether those values are human-desirable.

Goertzel (2016) proposes an alternative perspective that he calls ultimate value convergence:

> *Ultimate value convergence*: Nearly all superintelligent minds will converge to the same universal value system (paraphrased from Goertzel 2016, p.60).

Goertzel further proposes that the universal value system will be "centered around a few key values such as Joy, Growth, and Choice" (Goertzel 2016, p.60). However, the precise details of the universal value system are less important than the possibility that the value system could resemble human-desirable values. This creates a mechanism through which an AI that begins with any arbitrary human-undesirable value system could tend towards human-desirable values.

Goertzel does not insist that the ultimate values would necessarily be human-desirable. To the contrary, he states that "if there are convergent 'universal' values, they are likely sufficiently abstract to encompass many specific value systems that would be abhorrent to us according to our modern human values" (Goertzel 2016, p.60). Thus, ultimate value convergence does not guarantee that an AI would end up with human-desirable values. Instead, it increases the probability that an AI would end up with human-desirable values *if* the AI begins with human-undesirable values. Alternatively, *if* the AI begins with human-desirable values, then the ultimate value convergence theory could cause the AI to drift to human-undesirable values. Indeed, *if* the AI begins with human-desirable values, then more favorable results (from humanity's perspective) would accrue if the AI has goal-content integrity.

3 The ASI-PATH model

The ASI-PATH model was developed to model pathways to ASI catastrophe (Barrett and Baum 2016). ASI-PATH is a fault tree model, which means it is a graphical model with nodes that are connected by Boolean logic and point to some failure mode. For ASI-PATH, a failure mode is any event in which ASI causes global catastrophe. Fault tree models like ASI-PATH are used widely in risk analysis across a broad range of domains.

A core virtue of fault trees is that, by breaking catastrophe pathways into their constituent parts, they enable more detailed study of how failures can occur and how likely they are to occur. It is often easier to focus on one model node at a time instead of trying to study all potential failure modes simultaneously. Furthermore, the fault tree's logic structure creates a means of defining and quantifying model parameters and combining them into overall probability estimates. Indeed, the three points of the Bostrom-Goertzel disagreement (human evaluation of AI values, human creation of human-desirable AI values, and AI creation of human-desirable AI values) each map to one of the ASI-PATH parameters shown in Figure 1.

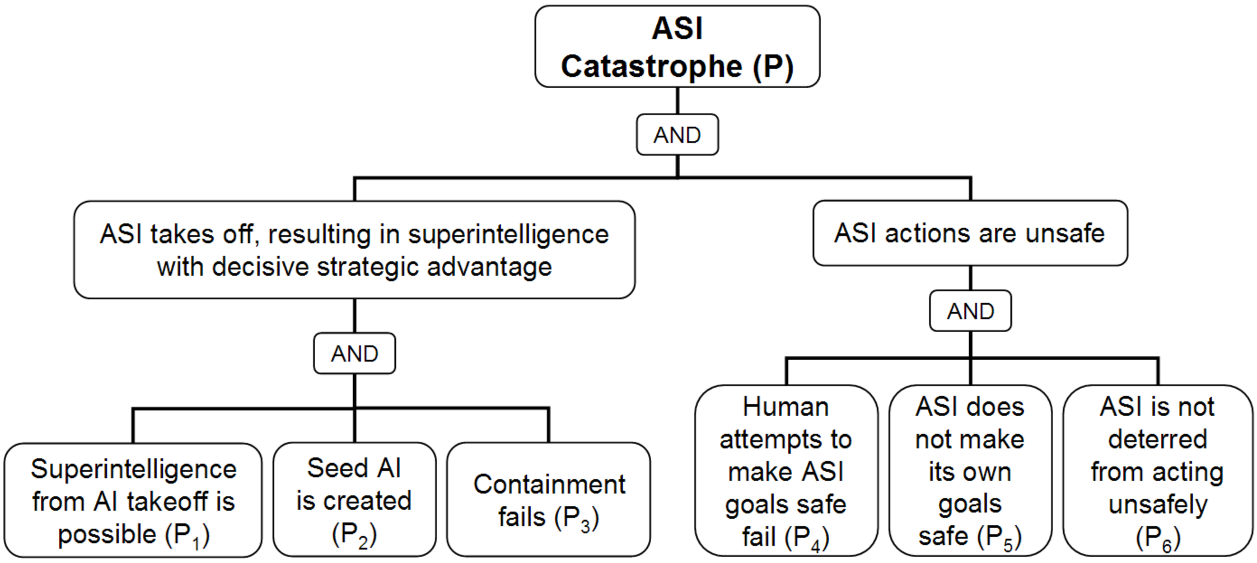
ASI Catastrophe (P)
AND
ASI takes off, resulting in superintelligence with decisive strategic advantage
ASI actions are unsafe
AND
AND
Superintelligence from AI takeoff is possible (P_1)
Seed AI is created (P_2)
Containment fails (P_3)
Human attempts to make ASI goals safe fail (P_4)
ASI does not make its own goals safe (P_5)
ASI is not deterred from acting unsafely (P_6)

Figure 1: ASI catastrophe fault tree. Adapted from Barrett and Baum (2017a).

In Figure 1, the top node is ASI catastrophe. The left branch covers events that lead to the ASI gaining "decisive strategic advantage", defined as "a level of technological and other advantages sufficient to enable it [the AI] to achieve complete world domination" (Bostrom, 2014, p. 78). The left branch models scenarios in which an initial "seed" AI undergoes recursive self-improvement and "takes off", becoming successively more and more intelligent until it becomes an ASI. P_1 is the probability that such an AI is possible in the first place. P_2 is the probability that a seed AI is created and undergoes recursive self-improvement. P_3 is the probability that the AI is contained from gaining decisive strategic advantage; the containment can occur at any point in the process from seed AI to ASI. Containment is any measure that prevents a seed AI from gaining decisive strategic advantage, either by limiting recursive self-improvement or by preventing ASI from gaining decisive strategic advantage. Containment includes confinement, in which the AI's ability to affect the rest of the world is restricted (Section 5), and enforcement, in which AI(s) prevent other AI(s) from gaining decisive strategic advantage.[2]

The left branch of Figure 1 covers events that could lead to the ASI taking actions that are "unsafe", which is defined as actions that would result in a major global

[2] Barrett and Baum (2017a, p. 400) define confinement as "restrictions built into the AI's hardware or software that limit the AI's ability to affect the rest of the world so that it does not gain decisive strategic advantage". This is slightly different than the Yampolskiy (2012) definition used in Section 5. This difference does not affect the overall argument of the present paper.

catastrophe. P_4 is the probability that humans will fail to make ASI goals safe. P_5 is the probability that the ASI will not make its own goals safe. Finally, P_6 is the probability that the ASI will not be deterred from acting unsafely by some other agent, potentially another AI. Because all the logic gates in Figure 1 are "AND", the probability of ASI catastrophe, P, is simply the product of the six component probabilities:

$$P = \prod_{n=1}^{6} P_n \qquad (1)$$

For convenience, we assume $\{P_1, P_2, P_6\} = 1$. These parameters are unrelated to the Bostrom-Goertzel disagreement as discussed in this paper. Instead, we focus on $\{P_3, P_4, P_5\}$, for which there is significant disagreement.

P_3 relates to the Bostrom-Goertzel disagreement about human evaluation of AI values (Section 2.1). In general, it should be easier to contain an AI earlier in the recursive self-improvement process because at that point it has less intelligence with which it could resist containment. Therefore, one factor in P_3 is the potential for human observers to determine early in the process that this particular AI should be contained. The easier it is for humans to evaluate AI values, the earlier in the process they should be able to notice which AIs should be contained, and therefore the more probable it is that containment will succeed. In other words, easier human evaluation of AI values means lower P_3.

P_4 relates to the Bostrom-Goertzel disagreement about human creation of human-desirable AI values (Section 2.2). Human-desirable values are very likely to be safe in the sense that they would avoid major global catastrophe. While one can imagine the possibility that somehow, deep down inside, humans actually prefer global catastrophe, and thus that an AI with human-desirable values would cause catastrophe, we will omit this possibility. Instead, we assume that an AI with human-desirable values would not cause catastrophe. Therefore, the easier it is for humans to create AIs with human-desirable values, the more probable it is that catastrophe would be avoided. In other words, easier human creation of AI with human-desirable values means lower P_4.

P_5 relates to the Bostrom-Goertzel disagreement about AI creation of human-desirable AI values (Section 2.3). We assume that the more likely it is that an AI would create of human-desirable values for itself, the more probable it is that catastrophe would be avoided. In other words, more likely AI creation of AI with human-desirable values means lower P_5.

For each of these three variables, we define two "expert belief" variables corresponding to Bostrom's and Goertzel's positions on the corresponding issue:

- P_{3B} is the value of P_3 that follows from Bostrom's position, the treacherous turn.
- P_{3G} is the value of P_3 that follows from Goertzel's position, the sordid stumble.
- P_{4B} is the value of P_4 that follows from Bostrom's position, the difficulty thesis.
- P_{4G} is the value of P_4 that follows from Goertzel's position, the weak difficulty thesis.
- P_{5B} is the value of P_5 that follows from Bostrom's position, goal-content integrity.
- P_{5G} is the value of P_5 that follows from Goertzel's position, ultimate value convergence.

Given estimates for each of the above "expert belief" variables, one can calculate P according to the formula:

$$P = \prod_{n=1}^{6} \left(W_{nB}P_{nB} + W_{nG}P_{nG} \right) \qquad (2)$$

In Equation 2, W is a weighting variable corresponding to how much weight one places on Bostrom's or Goertzel's position for a given variable. Thus, for example, W_{3B} is how much weight one places on Bostrom's position for P_3, i.e. how much one believes that an AI would conduct a treacherous turn. For simplicity, we assume $W_{nB} + W_{nG} = 1$ for $n = \{3, 4, 5\}$. This is to assume that for each of $\{P_3, P_4, P_5\}$, either Bostrom or Goertzel holds the correct position. This is a significant assumption: it could turn out to be the case that they are both mistaken. The assumption is made largely for analytical and expository convenience.

This much is easy. The hard part is quantifying each of the P and W variables in Equation 2. What follows is an attempt to specify how we would quantify these variables. We estimate the P variables by relating the arguments of Bostrom and Goertzel to the variables and taking into account any additional aspects of the variables. We aim to be faithful to Bostrom's and Goertzel's thinking. We estimate the W variables by making our own (tentative) judgments about the strength of Bostrom's and Goertzel's arguments as we currently see them. Thus, the P estimations aim to represent Bostrom's and Goertzel's thinking and the W estimations represent our own thinking. Later in the paper we also explore the implications of giving both experts' arguments equal weighting (i.e., $W_{nB} = W_{nG} = 0.5$ for each n) and of giving full weighting to exclusively one of the two experts.

We make no claims to having the perfect or final estimations of any of these parameters. To the contrary, we have low confidence in our current estimations, in the sense that we expect we would revise our estimations significantly in the face of new evidence and argument. But there is value in having some initial estimations to stimulate thinking on the matter. We thus present our estimations largely for sake of illustration and discussion. We invite interested readers to make their own.

3.1 P_3 and W_3: containment fails

The human evaluation of AI values is only one aspect of containment. Other aspects include takeoff speed (faster takeoff means less opportunity to contain AI during recursive self-improvement) and ASI containment (measures to prevent an ASI from gaining decisive strategic advantage). Therefore, the Bostrom-Goertzel disagreement about human evaluation of AI values should only produce a relatively small difference on P_3. Bostrom and Goertzel may well disagree on other aspects of P_3, but those are beyond the scope of this paper.

Bostrom's position, the treacherous turn, corresponds to a higher probability of containment failure and thus a higher value of P_3 relative to Goertzel's position, the sordid stumble. We propose a 10% difference in P_3 between Bostrom and Goertzel, i.e. P_{3B} - P_{3G} = 0.1. The absolute magnitude of P_{3B} and P_{3G} will depend on various case-specific details—for example, a seed AI launched on a powerful computer is more likely to have a fast takeoff and thus less likely to be contained. For simplicity, we will use P_{3B} = 0.6 and P_{3G} = 0.5, while noting that other values are also possible.

Regarding W_{3B} and W_{3G}, our current view is that the sordid stumble is significantly more plausible. We find it relevant that AIs are already capable of learning complex tasks like face recognition, yet such AIs are nowhere near capable of outwitting humans with a web of lies. Additionally, it strikes us as much more likely that an AI would exhibit human-undesirable behavior before it becomes able to deceive humans, and indeed long enough in advance to give humans plenty of time to contain the situation. Therefore, we estimate W_{3B} = 0.1 and W_{3G} = 0.9.

3.2 P_4 and W_4: humans fail to give AI safe goals

The Bostrom-Goertzel disagreement about human creation of human-desirable AI values is relevant to the challenge of humans giving AI safe goals. Therefore, the disagreement can yield large differences in P_4.

Bostrom's position, the difficulty thesis, corresponds to a higher probability of humans failing to give the AI safe goals and thus a higher value of P_4 relative to Goertzel's position, the weak difficulty thesis. The values of P_{4B} and P_{4G} will depend on various case-specific details, such as how hard humans try to give the AI safe goals. As representative estimates, we propose P_{4B} = 0.9 and P_{4G} = 0.4.

Regarding W_{4B} and W_{4G}, our current view is that the weak difficulty thesis is significantly more plausible. The fact that AIs are already capable of learning complex tasks like face recognition suggests that learning human values is not a massively intractable task. An AI would not please everyone all the time—this is impossible—but it could learn to have broadly human-desirable values and behave in broadly human-desirable ways. However, we still see potential for the complexities of human values to pose AI training challenges that go far beyond what exists for tasks like face recognition. Therefore, we estimate W_{4B} = 0.3 and W_{4G} = 0.7.

3.3 P_5 and W_5: AI fails to give itself safe goals

The Bostrom-Goertzel disagreement about AI creation of human-desirable AI values is relevant to the challenge of the AI giving itself safe goals. Therefore, the disagreement can yield large differences in P_5.

Bostrom's position, goal-content integrity, corresponds to a higher probability of the AI failing to give itself safe goals and thus a higher value of P_5 relative to Goertzel's position, ultimate value convergence. Indeed, an AI with perfect goal-content integrity will never change its goals. For ultimate value convergence, the key factor is the relation between ultimate values and human-desirable values; a weak relation suggests a high probability that the AI will end up with human-undesirable values. Taking these considerations into account, we propose P_{5B} = 0.95 and P_{5G} = 0.5.

Regarding W_{5B} and W_{5G}, our current view is that goal-content integrity is significantly more plausible. While it is easy to imagine that an AI would not have perfect goal-content integrity, due to a range of real-world complications, we nonetheless find it compelling that this would be a general tendency of AIs. In contrast, we see no reason to believe that AIs would all converge towards some universal set of values. To the contrary, we believe that an agent's values derive mainly from its cognitive architecture and its interaction with its environment; different architectures and interactions could lead to different values. Therefore, we estimate W_{5B} = 0.9 and W_{5G} = 0.1.

4 The probability of ASI catastrophe

Table 1 summarizes the various parameter estimates in Sections 3.1-3.3. Using these estimates, recalling the assumption $\{P_1, P_2, P_6\} = 1$, and following Equation 2 gives P = (0.1*0.6 + 0.9*0.5) * (0.3*0.9 + 0.7*0.4) * (0.9*0.95 + 0.1*0.5) ≈ 0.25. In other words, this set of parameter estimates implies an approximately 25% probability of ASI catastrophe. For comparison, giving equal weighting to Bostrom's and Goertzel's positions (i.e., setting each $W_B = W_G = 0.5$) yields P ≈ 0.26; using only Bostrom's arguments (i.e., setting each W_B = 1) yields P ≈ 0.51; and using only Goertzel's arguments (i.e., setting each W_G = 1) yields P = 0.1.

	P_B	P_G	W_B	W_G
3	0.6	0.5	0.1	0.9
4	0.9	0.4	0.3	0.7
5	0.95	0.5	0.9	0.1

Table 1: Summary of parameter estimates in Sections 3.1-3.3.

Catastrophe probabilities of 0.1 and 0.51 may diverge by a factor of 5, but they are both still extremely high. Even "just" a 0.1 chance of major catastrophe could warrant extensive government regulation and/or other risk management. Thus, however much Bostrom and Goertzel may disagree with each other, they would seem to agree that ASI constitutes a major risk.

However, an abundance of caveats is required. First, the assumption $\{P_1, P_2, P_6\} = 1$ was made without any justification. Any thoughtful estimates of these parameters would almost certainly be lower. Our intuition is that ASI from AI takeoff is likely to be possible, and ASI deterrence seems unlikely to occur, suggesting $\{P_1, P_6\} \approx 1$, but that the creation of seed AI is by no means guaranteed, suggesting $P_2 << 1$. This implies $P \approx 0.25$ is likely an overestimate.

Second, the assumption that the correct position was either Bostrom's or Goertzel's was also made without any justification. They could both be wrong, or the correct position could be some amalgam of both of their positions, or an amalgam of both of their positions plus other position(s). Bostrom and Goertzel are both leading thinkers about ASI, but there is no reason to believe that their range of thought necessarily corresponds to the breadth of potential plausible thought. To the contrary, the ASI topic remains sufficiently unexplored that it is likely that many other plausible positions can be formed. Accounting for these other positions could send P to virtually any value in [0, 1].

Third, the estimates in Table 1 were made with little effort, largely for illustration and discussion purposes. Many of these estimates could be significantly off, even by several orders of magnitude. Given the form of Equation 1, a single very low value for W_n*P_n would also make P very low. This further implies that $P \approx 0.25$ is likely an overestimate, potentially by several orders of magnitude.

Fourth, the estimates in Table 1 depend on a range of case-specific factors, including what other containment measures are used, how much effort humans put into giving the AI human-desirable values, and what cognitive architecture the AI has. Therefore, different seed AIs self-improving under different conditions would yield different values of P, potentially including much larger and much smaller values.

5 A practical application: AI confinement

A core motivation for analyzing ASI risk is to inform practical decisions aimed at reducing the risk. Risk analysis can help identify which actions would reduce the risk and by how much. Different assessments of the risk—such as from experts' differing viewpoints—can yield different results in terms of which actions would best reduce the risk. Given the differences observed in the viewpoints of Bostrom and Goertzel about ASI risk, it is possible that different practical recommendations could follow.

To illustrate this, we apply the above risk analysis to model the effects of decisions on a proposed ASI risk reduction measure known as AI confinement:

> *AI confinement*: The challenge of restricting an artificially intelligent entity to a confined environment from which it can't exchange information with the outside environment via legitimate or covert channels if such information exchange was not authorized by the confinement authority (Yampolskiy 2012, p.196).

AI confinement is a type of containment and thus relates directly to the P_3 (containment fails) variable in the ASI-PATH model (Figure 1). Stronger confinement makes it less likely that an AI takeoff would result in an ASI gaining decisive strategic advantage. Confinement might be achieved, for example, by disconnecting the AI from the internet and placing it in a Faraday cage.

Superficially, strong confinement would seem to reduce ASI risk by reducing P_3. However, strong confinement could increase ASI risk in other ways. In particular, by limiting interactions between the AI and the human populations, strong confinement could limit the AI's capability to learn human-desirable values, thereby increasing P_4 (failure of human attempts to make ASI goals safe). For comparison, AIs currently learn to recognize key characteristics of images (e.g., faces) by examining large data sets of images, often guided by human trainers to help the AI correctly identify image features. Similarly, an AI may be able to learn human-desirable values by observing large data sets of human decision-making, human ethical reflection, or other phenomena, and may further improve via the guidance of human trainers. Strong confinement could limit the potential for the AI to learn human-desirable values, thus increasing P_4.

Bostrom and Goertzel have expressed divergent views on confinement. Bostrom has favored strong confinement, even proposing a single international ASI project in which "the scientists involved would have to be physically isolated and prevented from communicating with the rest of the world for the duration of the project, except through a single carefully vetted communication channel (Bostrom 2014, p. 253)". Goertzel has explicitly criticized this proposal (Goertzel 2015, p.71-73) and instead argued that an open project would be safer, writing that "The more the AGI system is engaged with human minds and other AGI systems in the course of its self-modification, presumably the less likely it is to veer off in an undesired and unpredictable direction" (Goertzel and Pitt 2012, p.13). Each expert would seem to be emphasizing different factors in ASI risk: P_3 for Bostrom and P_4 for Goertzel.

The practical question here is how strong to make the confinement for an AI. Answering this question requires resolving the tradeoff between P_3 and P_4. This in turn requires knowing the size of P_3 and P_4 as a function of confinement strength. Estimating that function is beyond the scope of this paper. However, as an illustrative consideration, suppose that it is possible to have strong confinement while still giving the AI good access to human-desirable values. For example, perhaps a robust dataset of human decisions, ethical reflections, etc. could be included inside the confinement. In this case, the effect of strong confinement on P_4 may be small. Meanwhile, if there is no arrangement that could shrink the effect of confinement on P_3, such that this effect would be large, then perhaps strong confinement would be better. This and other practical ASI risk management questions could be pursued in future research.

6 Conclusion

Estimates of the risk of ASI catastrophe can depend heavily on which expert makes the estimate. A neutral observer should consider arguments and estimates from all available experts and any other sources of information. This paper analyzes ASI catastrophe risk using arguments from two experts, Nick Bostrom and Ben Goertzel. Applying their arguments to an ASI risk model, we calculate that their respective ASI risk estimates vary by a factor of five: $P \approx 0.51$ for Bostrom and $P = 0.1$ for Goertzel. Our estimates, combining both experts' arguments, is $P \approx 0.25$. Weighting both experts equally gave a similar result of $P \approx 0.26$. These numbers come with many caveats and should be used mainly for illustration and discussion purposes. More carefully considered estimates could easily be much closer to either 0 or 1.

These numbers are interesting, but they are not the only important part, or even the most important part, of this analysis. There is greater insight to be obtained from the details of the analysis than from the ensuing numbers. This is especially case for this analysis of ASI risk because the numbers are so tentative and the underlying analysis so comparatively rich.

This paper is just an initial attempt to use expert judgment to quantify ASI risk. Future research can and should do the following: examine Bostrom's and Goertzel's arguments in greater detail so as to inform the risk model's parameters; consider arguments and ideas from a wider range of experts; conduct formal expert surveys to elicit expert judgments of risk model parameters; explore different weighting techniques for aggregating across expert judgment, as well as circumstances in which weighted aggregation is inappropriate; conduct sensitivity analysis across spaces of possible parameter values, especially in the context of the evaluation of ASI risk management decision options; and do all of this for a wider range of model parameters, including $\{P_1, P_2, P_6\}$ as well as more detailed components of $\{P_3, P_4, P_5\}$, such as modeled in Barrett and Baum (2017a; 2017b). Future research can also explore the effect on overall ASI risk when multiple ASI systems are launched: perhaps some would be riskier than others, and it may be important to avoid catastrophe from all of them.

One overarching message of this paper is that more detailed and rigorous analysis of ASI risk can be achieved when the risk is broken into constituent parts and modeled, such as in Figure 1. Each component of ASI risk raises a whole host of interesting and important details that are worthy of scrutiny and debate. Likewise, aggregate risk estimates are better informed and generally more reliable when they are made from detailed models. To be sure, it is possible for models to be too detailed, burdening experts and analysts with excessive minutiae. However, given the simplicity of the risk models at this early stage of ASI risk analysis, we believe that, at this time, more detail is better.

A final point is that the size of ASI risk depends on many case-specific factors that in turn depend on many human actions. This means that the interested human actor has a range of opportunities available for reducing the probability of ASI catastrophe. Risk modeling is an important step towards identifying which opportunities are most effective at reducing the risk. ASI catastrophe is by no means a foregone conclusion. The ultimate outcome may well be in our hands.

7 Acknowledgement

We thank Ben Goertzel, Miles Brundage, Kaj Sotala, Steve Omohundro, Allan Dafoe, Stuart Armstrong, Ryan Carey, Nell Watson, and Matthijs Maas for helpful comments on an earlier draft. Any remaining errors are the authors' alone. Work for this paper is funded by Future of Life Institute grant 2015-143911. The views in this paper are those of the authors and do not necessarily reflect the views of the Global Catastrophic Risk Institute or the Future of Life Institute.

8 References

[1] Armstrong S, Sotala K (2012). How we're predicting AI—or failing to. In Romportl J, Ircing P, Zackova E, Polak M, Schuster R (eds), Beyond AI: Artificial Dreams. Pilsen, Czech Republic: University of West Bohemia, pp. 52-75.

[2] Armstrong S, Sotala K, Ó hÉigeartaigh SS (2014). The errors, insights and lessons of famous AI predictions – and what they mean for the future. Journal of Experimental & Theoretical Artificial Intelligence 26(3), 317-342.

[3] Barrett AM, Baum SD (2017a). A model of pathways to artificial superintelligence catastrophe for risk and decision analysis. Journal of Experimental & Theoretical Artificial Intelligence 29(2), 397-414.

[4] Barrett AM, Baum SD (2017b). Risk analysis and risk management for the artificial superintelligence research and development process. In Callaghan V, Miller J, Yampolskiy R, Armstrong S (eds), The Technological Singularity: Managing the Journey. Berlin: Springer, pp. 127-140.

[5] Baum SD, B Goertzel, TG Goertzel (2011). How long until human-level AI? Results from an expert assessment. Technological Forecasting & Social Change 78(1), 185-195.

[6] Bostrom N (2014). Superintelligence: Paths, Dangers, Strategies. Oxford: Oxford University Press.

[7] Goertzel B (2015). Superintelligence: Fears, promises and potentials. Journal of Evolution and Technology 25(2), 55-87.

[8] Goertzel B (2016). Infusing advanced AGIs with human-like value systems: Two theses. Journal of Evolution and Technology 26(1), 50-72.

[9] Goertzel B, Pitt J (2012). Nine ways to bias open-source AGI toward friendliness. Journal of Evolution and Technology 22(1), 116-131.

[10] Müller VC, Bostrom N (2014). Future progress in artificial intelligence: A survey of expert opinion. In

Müller VC (ed), Fundamental Issues of Artificial Intelligence. Berlin: Springer, pp. 555-572.

[11] Oreskes N (2004). The scientific consensus on climate change. Science 306(5702), 1686.

[12] Yampolskiy R (2012). Leakproofing the Singularity: Artificial intelligence confinement problem. Journal of Consciousness Studies 19(1-2), 194-214.