⏵ FHI TECHNICAL REPORT ⏴

Racing to the Precipice: a Model of Artificial Intelligence Development

Stuart Armstrong
Nick Bostrom
Carl Shulman

Technical Report #2013-1

Cite as:

Armstrong, S. & Bostrom, N. & Shulman, C. (2013): "Racing to the precipice: a model of artificial intelligence development", *Technical Report* #2013-1, Future of Humanity Institute, Oxford University: pp. 1-8.

The views expressed herein are those of the author(s) and do not necessarily reflect the views of the Future of Humanity Institute.

Racing to the precipice: a model of artificial intelligence development

Stuart Armstrong Nick Bostrom Carl Shulman

October 2013

Abstract

This paper presents a simple model of an AI arms race, where several development teams race to build the first AI. Under the assumption that the first AI will be very powerful and transformative, each team is incentivised to finish first – by skimping on safety precautions if need be. This paper presents the Nash equilibrium of this process, where each team takes the correct amount of safety precautions in the arms race. Having extra development teams and extra enmity between teams can increase the danger of an AI-disaster, especially if risk taking is more important than skill in developing the AI. Surprisingly, information also increases the risks: the more teams know about each others' capabilities (and about their own), the more the danger increases.

1 Introduction

This paper presents a simplified model for analysing technology races. The model was designed initially for races to construct artificial intelligences (AIs). But it can be applied to other similar races or competitions, especially technological races where there is a large advantage to reaching the goal first.

There are arguments that the first true AIs are likely to be extremely powerful machines [Goo65, Cha10], but that they could end up being dangerous [Omo08, Yud08] if not carefully controlled [ASB12]. The purpose of various research projects such as 'friendly AI' [MS12] is to design safety precautions ensuring the creation of an AI with human-compatible values.

In this paper, we won't go that far. We will simply assume that there is a definite probability of an AI-related disaster, and that the probability goes up the more the AI development team skimps on precautions. This paper will present a model of an 'arms race' between AI development teams, and analyse what factors increase and decrease the probability of such an AI-disaster.

Several factors contribute to increasing the danger: if building the AI depends more on risk-taking than on skill, for instance. Reducing the enmity between AI development teams helps, however, as does reducing the number of teams.

But surprisingly, extra information can exacerbate the danger. The danger is minimised when each team is ignorant of the AI-building capabilities of every team – including their own capabilities. This is an interesting example of an information hazard [Bos11].

2 The model

In this model, there are n different teams, competing to build the first proper AI. Each team has a certain level of AI-building capability c, and can choose to take a certain level of safety precautions s, varying between $s = 1$ (total safety) and $s = 0$ (complete lack of precautions). Then each team gets a score by subtracting their safety from their capability ($c - s$), and the team with the highest score 'wins', by building their AI first.

Once that AI is built, we need to check for the success of the project: That will depend on the safety level of the winning team[1]. With probability s, they have a successful and profitable AI project. With probability $1 - s$, they have caused an AI-disaster.

We'll assume that each team maximises expected utility, and we can renormalise their utilities to give utility 1 for a successful AI project, and 0 for an AI-disaster.

We will further assume that each team has the same level of enmity e towards each other team. This means that they will have a utility of $1 - e$ if another team builds a successful AI before they do. This varies between $e = 0$ (they are indifferent to who builds an AI) and $e = 1$ (another team building a successful AI is just as bad as an AI-disaster). It would be possible to have enmities above 1 (this would correspond to teams that hate each other so much that an AI-disaster is preferable to the triumph of the other team), but that is beyond the scope of the current paper.

Further, we'll assume that each team's capability is drawn independently and uniformly on the interval $[0, \mu]$, for a single given μ. A high μ corresponds to a situations where capability is the dominant factor in AI development: one can achieve very little extra by skimping on precautions. Conversely, for low μ, incurring increased risk can make one much more likely to build the first AI.

We'll assume that the teams choose their safety based on Nash equilibrium considerations (this could happen in practice if, for instance, each team reduced their safety a little bit, then a little bit more, then a little bit more, in response to the other teams reducing their own safety). We'll then calculate the Nash equilibria in three information scenarios: when teams don't know anyone's capabilities, when they know their own capabilities, and when they know everyone's capabilities[2].

2.1 No information

In this situation, every team is ignorant of their own or any other team's capabilities[3]. Each team will thus choose the level of safety the think is best. In

[1] And only the winning team – if another team gets a disastrous AI first by taking lower precautions, they will 'won' the race to build the first AI.

[2] Of course, the model can be refined in various ways. One could make capacity information uncertain and fuzzy, one could have different levels of enmity between different teams, one could incorporate uncertainty about the safety levels and the ultimate outcomes, and so on. Or one could have a dynamic process to determine the outcome, rather than rushing straight to the Nash equilibrium. But the simple model is enough to gain useful insights.

[3] It may seem unusual for teams to not know their own capabilities in the real world. However, this is close to the situation we find ourselves with current AI research: people and organisations have a pretty clear idea of what resources and knowledge they have, but don't know how hard AI is or what routes are most likely to lead there. They are thus effectively ignorant of their own AI-building capabilities.

a symmetric Nash equilibrium, this will be a single number s. Since all teams draw their capabilities from independent identically distributed random variables, and have the same level of safety, they each have the same probability of winning, namely $1/n$.

Upon changing s to $s - \epsilon$, a team's chance of winning increases by

$$\epsilon \frac{1}{\mu} + O(\epsilon).$$

The utility gain from winning rather than letting another win is the safety level s times the enmity e. So the extra expected utility from winning is

$$\epsilon \frac{se}{\mu} + O(\epsilon).$$

But the decrease in safety brings in extra risks, were the team to win anyway. This extra risk is simply ϵ time the probability of winning:

$$\frac{\epsilon}{n} + O(\epsilon).$$

All other effects are second order in ϵ. Setting the change to be zero (to first order) in ϵ, the general solution is thus

$$s = \begin{cases} \frac{\mu}{en} & \mu < en \\ 1 & \mu \geq en. \end{cases}$$

The total probability of AI-disaster is $1 - s$, namely:

$$\begin{cases} 1 - \frac{\mu}{en} & \mu < en \\ 0 & \mu \geq en. \end{cases}$$

2.2 Private information

In this scenario, each team knows its own capabilities, but not those of the other teams. Each team will choose a safety level $s(x)$, dependent on their own capabilities x. In a symmetric Nash equilibrium, each team chooses the same function s. We may assume $x - s(x)$ is a function increasing in x[4]. In that case, the team with the highest x, the highest capabilities, always wins.

Given x, the probability of a team winning is equal to the probability of all other teams having lower capability: this is $\left(\frac{x}{\mu}\right)^{n-1}$. Then changing $s(x)$ to $s(x) - \epsilon$, the team increases its probability of winning by:

$$\epsilon \frac{(n-1)x^{n-2}}{(1 - s'(x))\mu^{n-1}} + O(\epsilon).$$

As in the previous case, the expected extra utility is se times this:

$$\epsilon \frac{se(n-1)x^{n-2}}{(1 - s'(x))\mu^{n-1}} + O(\epsilon).$$

[4] If makes no sense that a team with higher capability would have a lower chance of winning (if so, they would voluntarily destroy part of their capability).

The loss in expected utility coming from winning at a lower safety level is:

$$\epsilon \left(\frac{x}{\mu}\right)^{n-1} + O(\epsilon).$$

Solving these equations gives:

$$s(x) = \begin{cases} \frac{x}{en-e+1} & x < en - e + 1 \\ 1 & x \geq en - e + 1. \end{cases}$$

The total probability of a AI-disaster is calculated by integrating, across all values of x in $[0, \mu]$, the risk level $1 - s(x)$ times the probability that the winning team will have capability x:

$$\int_0^\mu (1 - s(x)) \frac{nx^{n-1}}{\mu^n} dx = \begin{cases} 1 - \frac{\mu n}{(n+1)(ne-e+1)} & \mu < en - e + 1 \\ \frac{(en-e+1)^n}{(n+1)\mu^n} & \mu \geq en - e + 1. \end{cases}$$

2.3 Public information

In this scenario, every team knows the capabilities of every other team. This scenario is analysed somewhat differently that the other. Let Δ be the difference between the capability of the top team and the second ranked team. The top team always wins, but its safety level s_{top} is determined by Δ. When $s_{top}(\Delta) = \Delta/e$, it is not in the interest of the second team to decrease its safety to compete with the top team. The gain from winning does not compensate for the extra risks run.

Thus the safety of the top team will be $s_{top} = \Delta/e$ if $\Delta/e < 1$ and 1 otherwise. The total probability AI-disaster is calculated by integrating, across all values of Δ in $[0, \mu]$, the risk level $1 - s_{top}(\Delta) = 1 - \Delta/e$ times the probability that the difference between the top two teams is Δ:

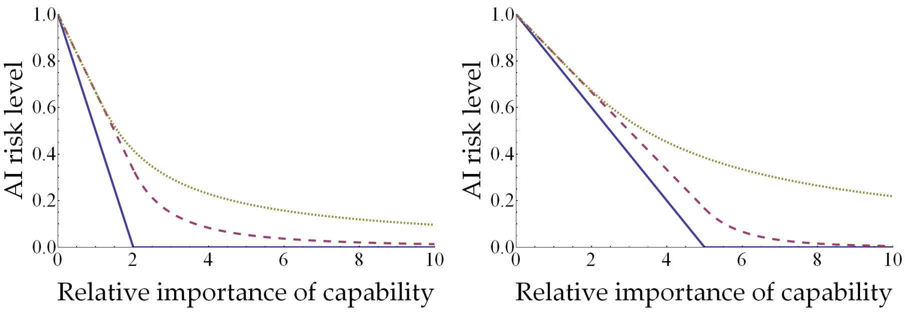
$$\int_0^\mu (1 - s_{top}(\Delta)) \frac{n(\mu - \Delta)^{n-1}}{\mu^n} d\Delta = \begin{cases} 1 - \frac{\mu}{e(n+1)} & \mu < e \\ \frac{(\mu-e)^{n+1}}{e(n+1)\mu^n} - \frac{\mu}{e(n+1)} + 1 & \mu \geq e. \end{cases}$$

3 Factors determining risk

3.1 Capability vs risk

Intuitively it is clear that increasing the importance of capability must decrease overall risk. One is less inclined to skimp on safety precautions if can only get a small advantage from doing so. This intuition is born out by the results: in every situation, an increase of the importance of capability (an increase in μ) reduces the risk. This is illustrated in figures 1a and 1b, as well as all subsequent figures, demonstrating that the probability of AI-disaster are always decreasing in μ. Indeed, around $\mu = 0$ (i.e. when capability is nearly irrelevant to producing the first AI), the only Nash equilibrium is to take no safety precautions at all.

In terms of intervention, there is little we can do about the relative importance of capability, since this is largely determined by the technology. Our best bet might be at present to direct research into approaches where there is little return to risk-taking, prioritising some technological paths over others.



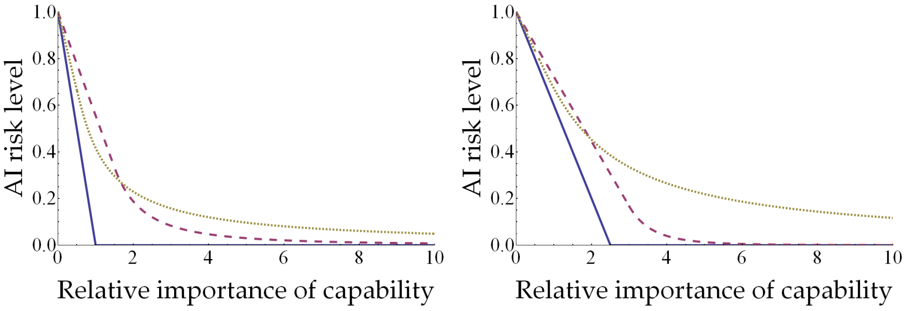
(a) Two teams

(b) Five teams

Figure 1: Risk of dangerous AI arms race for two and five teams, at enmity 1, plotted against relative importance of capability. Three information-level scenarios: no capability information (full), private capability information (dashes), and full capability information (dots).

3.2 Compatible goals

It is also intuitively clear that reducing enmity should reduce the risk of AI-disaster. When competition gets less fierce, teams would be willing to take less risks. This intuition is also born out in our model, as decreases in e always reduce the risk. For illustration, contrast the graphs 2a and 2b, where the enmity is 0.5, with the previous graphs 1a and 1b where the enmity was 1.



(a) Two teams

(b) Five teams

Figure 2: Risk of dangerous AI arms race for two and five teams, at enmity 0.5, plotted against relative importance of capability. Three information-level scenarios: no capability information (full), private capability information (dashes), and full capability information (dots).

Enmity is something that we can work on by, for instance, building trust between nations and groups, sharing technologies or discoveries, merging into joint projects or agreeing to common aims. With this extra coordination, we could also consider agreements to allow the teams to move away from the Nash equilibrium, thus avoiding a race to the bottom when the situation is particularly dangerous (such as low capability μ). Friendly teams make friendly AIs.

3.3 The curse of too much information

Contrasting figure 1a with figure 2a (or figure 1a with figure 2b), one notices a curious pattern. The no-information case is always safer than the other two cases, but the relative safety of private information or common information depends on the degree on enmity.

It is always better if none of the teams have any idea about anyone's capability. For maximal enmity ($e = 1$), the private information scenario is similarly safer than the common information one. But for lower e, this does not hold: for low μ and low e, the public information scenario can be safer than the private one. Asymptotically, though, the private information case is of order of $1/\mu^n$, while the public information case is of order $1/\mu$.

The reasons for this is that it is only worth taking risks in the private information if one's capability is low. So to get a high risk, one needs a winner with low capability, which is equivalent to having all teams at low capability. The probability of having a single team at low capability diminishes inversely with μ. Hence the probability of having all teams at low capability diminishes as $1/\mu^n$. In the public information case, the winner will take risks if the second ranked team is close to its own capability. Being close to a specific other team is inversely proportional to μ, and hence the probability of this diminishes as $1/\mu$.

Our ability to manipulate the information levels of the teams or teams is likely to be somewhat limited, though we can encourage them to share more (in the low capability, low enmity situations) or alternatively to lock up their private knowledge[5] (in situation of higher capability or enmity).

3.4 The number of competitors

Finally, we can consider what happens when more teams are involved. Intuitively, competition might spur a race to the bottom if there are too many teams, and the model bears this out: in both the no-information and public information cases, adding extra teams strictly increases the dangers.

The private information case is more complicated. At low capability, adding more teams will certainly worsen the situation, but at higher capability, the effect is reversed[6]. This can be seen in figure 3, which plots the private information risk curves for two teams and five teams.

In terms of leverage, the most helpful intervention we could undertake to reduce the number of teams is to encourage groups to merge. Dividing teams would only be useful in very specific situations, and would need some method of ensuring that the divided teams did not recall each other's capabilities.

4 Conclusion

We've presented a model of AI arms race (which can be extended to other types of arms races and competitions) in which teams of AI designers can get ahead by skimping on safety. If the race takes some time there is a persistent

[5] Such secrecy can interfere with trust building, though, making it hard to reach agreements between teams if such agreement is needed.

[6] This is because only the teams with low capability take risks in cases of private information, and the more teams there are, the less likely it is that the winner will be low capability.

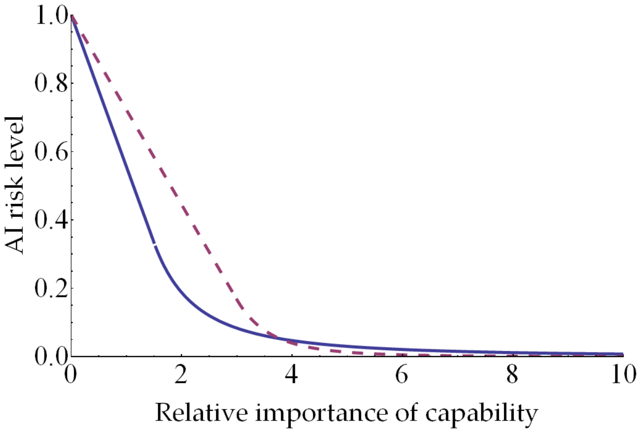


Figure 3: Risk of dangerous AI arms race for private information teams, at enmity 0.5, plotted against relative importance of capability. The graph for two teams is full, the graph for five teams is dashed.

ratchet effect (or race to the bottom) in which each team reduces their safety precautions slightly, until a Nash equilibrium is reached– possibly one of very high danger.

Several factors influence on the probability of AI disaster, three in an intuitive way, and one counter-intuitively.

Intuitively, if capability is much more important than the level of security precautions taken, the overall outcome is safer. Similarly, reduced enmity between the teams produces better outcomes, as does reducing the number of teams (in most situations).

Counter-intuitively, increasing the information available to all the teams (about their own capability or progress towards AI, or that of the other teams) increases the risk. This is a special case of an information hazard [Bos11]: we'd be better off not knowing.

This model and variants of if may be of use in planning and regulating the development of AIs and other disruptive technological innovations.

5 Acknowledgments

We are grateful to Anders Sandberg and Seán S. ÓhÉigeartaigh for helpful discussions.

References

[ASB12] Stuart Armstrong, Anders Sandberg, and Nick Bostrom. Thinking inside the box: Controlling and using an oracle ai. *Minds and Machines*, 22:299–324, 2012.

[Bos11] Nick Bostrom. Information hazards: A typology of potential harms from knowledge. *Review of Contemporary Philosophy*, pages 44–79, 2011.

[Cha10] David Chalmers. The singularity: A philosophical analysis. *Journal of Consciousness Studies*, 17:7–65, 2010.

[Goo65] I.J. Good. Speculations concerning the first ultraintelligent machine. *Advances in Computers*, 6, 1965.

[MS12] Luke Muehlhauser and Anna Salamon. Intelligence explosion: Evidence and import. In A. Eden, J. Søraker, J.H. Morr, and E. Steinhart, editors, *The singularity hypothesis: A scientific and philosophical assessment*. Berlin: Springer, 2012.

[Omo08] Stephen M. Omohundro. The basic ai drives. *Frontiers in Artificial Intelligence and applications*, 171:483–492, 2008.

[Yud08] Eliezer Yudkowsky. Artificial intelligence as a positive and negative factor in global risk. In Nick Bostrom and Milan M. Ćirković, editors, *Global catastrophic risks*, pages 308–345, New York, 2008. Oxford University Press.