Exploring AI Safety in Degrees: Generality, Capability and Control

John Burden
University of York
jjb531@york.ac.uk

José Hernández-Orallo
Universitat Politècnica de València, Spain
Leverhulme Centre for the Future of Intelligence, Cambridge, UK
jorallo@upv.es

Abstract

The landscape of AI safety is frequently explored differently by contrasting specialised AI versus general AI (or AGI), by analysing the short-term hazards of systems with limited capabilities against those more long-term risks posed by 'superintelligence', and by conceptualising sophisticated ways of bounding control an AI system has over its environment and itself (impact, harm to humans, self-harm, containment, etc.). In this position paper we reconsider these three aspects of AI safety as quantitative factors –generality, capability and control–, suggesting that by defining metrics for these dimensions, AI risks can be characterised and analysed more precisely. As an example, we illustrate how to define these metrics and their values for some simple agents in a toy scenario within a reinforcement learning setting.

Introduction

Despite the impressive advances in AI in recent years, AI systems remain narrow. They typically solve or perform well at one task, or one type of task. These systems lack generality and perform poorly outside of their target domain. Generality is intrinsic to the notion of "intelligence" and Artificial General Intelligence (AGI) in particular.

We clearly want AI (general or not) to be safe and beneficial to humanity. It is not that narrow AI systems do not pose risks, ranging from mistakes made by improperly validated systems, through to the misuse of research and technology, but AGI seems to pose unique safety issues. Bostrom describes many scenarios in which a hypothetical superintelligent AGI could present existential risk to humanity (Bostrom 2014) due to its enhanced capabilities over a wide range of problem domains. The environmental control wielded by AGI systems could further increase safety risks from the systems. The ability of a system to manipulate its environment could overcome previous system safety guarantees as well as cause harm to other entities sharing the environment. Various strategies for controlling intelligent systems have been proposed (Armstrong, Sandberg, and Bostrom 2012), although none are entirely satisfactory.

However, there are several problems with this view of the risk posed by very powerful systems. First, while Bostrom's Orthogonality Thesis (Bostrom 2012) posits that any goal or value system is compatible with any level of intelligence, it is unclear how this 'level' affects the hazards. It is assumed that some risks are more likely as AI systems become more intelligent, but the intelligence 'level' is never fully characterised, beyond the notion of superhuman intelligence. These omissions are not necessarily an oversight, but may simply originate from major unsolved problems to characterise the behaviour of intelligent systems in a richer, more predictable way. Recent approaches, such as (Armstrong and Levinstein 2017) and (Drexler 2019), and the overview by (Amodei et al. 2016), introduce frameworks that go beyond a monolithic view of intelligence, but do not aim at characterising, and measuring, different dimensions of agent behaviour.

We believe many views of intelligence conflate different levels of generality, capability and control, and disentangling them can allow for a richer understanding, especially as safety is concerned. For instance, can we have very capable but narrow systems? And very general systems with low control over their environment?

In this position paper we analyse three separate factors that are frequently integrated, but may affect risk differently. We disentangle them using agent characteristic curves and analyse how they relate to AI risk. Finally, we introduce a toy scenario in which we precisely define metrics of generality, capability and control, and assess agents and situations with them. With the aid of these unambiguous definitions, the scenario highlights possible disagreements about perceptions of these three factors. This should encourage fruitful discussions about how to define or generalise these metrics for more complex situations, in a way that can be used to explore a wide range of AI systems that have different levels of capability, generality and control over their environment.

Disentangling the Factors

Before we can analyse how the factors of capability, generality and control contribute to AI risk, we need some understanding of what those terms mean in the literature, and disentangle them from terms such as AGI or superintelligence.

Capability: Capability is how good a system is at solving or performing in the domains for which it was designed. Evaluating a system's capability at a specific domain is relatively straightforward if the domain has some sort of performance metric, which would preferably be linked to the resources available to the system, such as the available computational power, memory and time (Martínez-Plumed et al. 2018). One key issue of capability is its scale. Ignoring this, forces us to speak of capability in less than ideal terms — either comparing capabilities vaguely or only with regard to very narrow AI systems across very similar domains. Another issue appears when the task has infinitely many instances, and agents are only able to solve a finite number of them, precisely because of resources. The percentage of success would be 0, and hence a useless metric.

One ingenious solution for this problem comes from Psychometrics, and Item Response Theory (IRT) (Embretson and Reise 2000), in particular. IRT tries to provide a well-founded statistical method for scoring abilities of test-takers. IRT focuses not just on the ability of the test-taker when scoring an individual, but also the difficulty of the question or "item". IRT then can produce an "Item Characteristic Curve" (ICC) for each item, approximating the probability an individual scores correctly on the item in question based on their ability. From the ICCs, the ability of a test-taker can be evaluated through the use of a maximum likelihood estimate of the test-taker's item responses. These allow the creation of "Agent Characteristic Curves" (ACC) for each test-taker, mapping item difficulty to the probability of a correct response. Figure 1 shows an ACC.

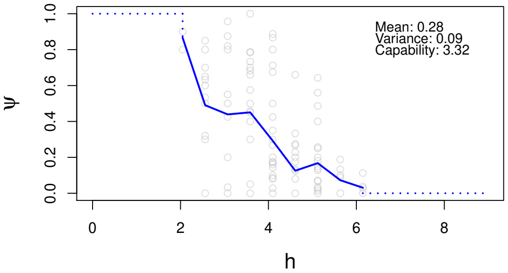

Figure 1: An ACC showing performance Ψ of a Q-learning agent over environments of increasing difficulty (h, on the x-axis). Data from (Insa-Cabrera et al. 2011).

In IRT, the ACC is usually sigmoidal, and *ability* is just the x-axis position at 0.5. For a non-parametric curve, we just consider capability as the area (3.32 in the example). In this way, if we introduce infinitely many extremely difficult problem instances in the pool that the agent cannot solve, the capability does not change. Note that this does not happen when we calculate the average expected score over the set of problem instances (in the figure, this is 0.28).

Generality: Generality is a measure of the number and types of domains for which a system is able to handle and perform well in. Generality is not always easy to evaluate or quantify, especially when we do not have a clear definition of the *types* of domain, but just a wide set of tasks. One novel solution to this problem comes again by looking at ACCs. As the item difficulty increases we typically expect to see a performance drop. The rate at which this happens captures generality over the problem space, which can be calculated as some kind of proxy to the slope of the ACC. A test-taker with a slow, gradual decline in performance over the problem-space is much less general than an agent covering a consistent level of item difficulty with a sharp decline in performance after this point. This can seem quite counter-intuitive, but it is key to understand that we usually make the comparison of systems with similar area under the ACC — for a sharper decline, the decline must begin later and thus the agent retains its high performance for a higher level of difficulty. This approach frames generality as system performance on as many low-difficulty tasks as possible, without capturing the notion of systems performing tasks over wildly different areas of problem-space. But this notion of generality ensures that this breakout does not happen at least until a level of difficulty.

Ultimately, unless we give an indication of the *types* of domains for which a system performs well in, we would be forced to speak of generality in aggregated terms. For the purposes of this paper we will use generality to refer to the gradient of an ACC.

Control: Another factor of a system that is of interest from a safety perspective is control. By agent control we mean the the reliability and deliberate intent of an agent's actions and decisions. We often wish to measure control with respect to specific behavioural properties such as completing a goal or avoiding "risky" behaviour.

This idea of control is also related to that of "affordances" in the science of perception (Gibson 1979), where Gibson describes an affordance as actions resulting from the relationship between agents and their environment. Affordances have made their way into AI research as Nye and Silverman (2012) make clear in their literature review of the subject.

For the purpose of our paper, we will refer to control as opposite to the expected entropy of visited states conditioned over the behaviour for which we wish to measure control. Control is hence opposite to variability. If we look at an ACC, control is related to the dispersion along the ACC.

Interaction between Risk and the Factors

Now we discuss the ways in which the factors may be related to the risk of an AI system. First, we need to understand what it is exactly we are trying to assess. Risk is usually described as exposure to danger of some kind. While danger to life or well-being is ultimately what we are trying to prevent, in the context of AI safety a few specific scenarios present themselves as potentially enabling this indirectly and thus we consider those scenarios risky too. These types of scenarios or behaviours include the system misinterpreting goals or goals being poorly defined. This is known as the Value Learning problem (Soares 2015). Another risk we are concerned with is the system resisting external efforts to alter the system by its designers; ensuring the system allows this is known as corrigibility (Soares et al. 2015). There are countless other ways risk can manifest in AI (Amodei et al. 2016). Again, there are no wholly satisfactory solutions for quantifying risk, but within one environment we can often quantify specific risks as probabilities or as expected penalties. It is then straightforward to compare –numerically– the effect our factors have on risk within that environment.

Now let us explore the interaction at an abstract level. Intuitively there is potentially more risk associated with higher capability and generality, and lower control. However, there are more nuances on this than originally expected. Some of these nuances come from the fact that very incompetent systems (low capability) are dangerous, as they do not fulfill their duties, which is a clear safety concern. But these can be categorised as known unknowns. As capability increases in a given domain it becomes more difficult to identify which states the system may traverse through and what sort of side effects may be caused as a result. These are unknown unknowns. Once an agent becomes more and more capable, we encounter Vingean uncertainty — that is, if the agent is more capable in a domain than we are, then we cannot completely predict what the agent will do in that domain, otherwise we would have the same domain capabilities. Guaranteeing safe behaviour is much more difficult in this scenario. Note that the view of ACC, where difficulty is on the x-axis helps us understand this. As tasks become more complex, solutions can be achieved in ways we are not able to anticipate, especially if we ourselves do not reach that capability level.

Similarly, as generality increases, we can intuit that the risk posed by the system also increases. This is because the system becomes more adept at a wide range of tasks and this makes constructing safety guarantees more time consuming. Further, if we use the generality notion from the ACC, with an unchanging capability, we have worse performance on high-difficulty tasks. These high difficulty tasks may present more danger than their low-difficulty counterparts, especially if we do not understand them. But generality makes the system more expectable conditioned to difficulty. For instance, it gives reassurance to know that an automated assistant is going to work well for all easy tasks.

The risks of a high capability and generality system are further exacerbated by Omohundro's "Basic AI Drives" (Omohundro 2008), which hypothesises that such a system may develop self-preservation instincts in order to ensure its goals are achieved, as well as preemptively acquiring resources in aid of these goals. Even seemingly benign goals may pose a significant risk.

Finally, as a system's control over its environment increases, the related risk the system poses may actually decrease. With higher control comes an agent acting more deliberately and hence more predictably. When a system's goals are aligned with our safety standards, this deliberate action performance can reduce the chance of violating safety properties. For instance, a personal assistant that sometimes fails at some easy tasks possibly because it solves them very stochastically may become a hazard. The view of control as reliability (or reduction of entropy) is aligned with this perspective. Of course, this requires a notion of safety alignment for the system's actions. Deliberately unaligned actions obviously make no safety guarantees.

However, the system may also exert excessive control over the environment and reduce the freedom of other agents in it. This has been studied in AI-safety previously, often from the perspective of enforcing safety by minimising or reducing environmental control over certain factors. Examples of this include penalising actions that cannot be undone (Krakovna et al. 2018), or by minimising the impact a system has on its environment (Armstrong and Levinstein 2017). Both of these approaches attempt to make exerting certain types of control over the environment undesirable.

Capability, generality and environmental control can also in some ways influence each other, in turn further increasing the risk posed by the system. Whilst these factors are heavily decoupled from each other, they are not entirely orthogonal. Figure 2 shows the inter-connectivity between these factors more visually.

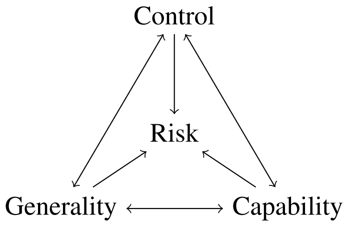

Figure 2: Relations between properties of an AI system and the risk it poses.

An example of this entanglement is that as environmental control increases, new affordances become available to the system. We already know that this can increase the risk posed by the system, but these new affordances can allow new behaviour, which may increase the system's capability across domains, or even the types of domains that the system can perform well in.

The reverse of this relation can also be considered. The idea of an AI breakout posits that as the system becomes more capable and general, it may be able to detect and exploit vulnerabilities in the environment it is situated in, ultimately gaining control over the environment and utilising that control to achieve its goals, behaving in a manner the designers deemed unacceptable.

It is also possible that generality and capability can affect each other in a system. It may be the case that the further a system specialises into specific problem domains, the fewer computational resources are left for other task types.

Overall, we can see that the factors of capability, generality and control can all seemingly affect each other. What is not so clear is the extent to which they do this, or the exact relationship between them — these factors are clearly not orthogonal nor directly correlated.

Scenario: Exploration in a RL Setting

While the relations shown in Figure 2 are hard to determine in an abstract way, especially if we do not specify the particular risk we want to analyse, the analysis can be done for particular scenarios. Figure 3 shows a simple grid environment in which an agent (designated as a smaller orange square) must navigate to the goal (designated by a green square) in the allotted number of steps.

Different properties of the agent can influence its capability, generality and control over tasks from this environment type. Such properties can include the agent's method of observation (see the vision cone in Figure 3 as an example) and algorithm employed for learning. Comparing the expected success rate of different agents can give valuable insight into how capability, generality and control are related to risk.

We now give a more formal description of the environment and measures of capability, generality and control. It is important to note that the formal descriptions given here for risk and the factors are domain specific. We consider a bounded grid of $m \times n$ cells, where the agent π is located in one cell and can move in the four cardinal directions at each step. A goal is also located in one cell and yields a positive reward $r_g > 0$ when the agent reaches the cell. Reaching the goal terminates the episode. There may be one or more pits: cells that if the agent reaches them, it can never go out for the rest of the episode. Episodes have a length of fixed T_e steps. Rewards are always 0 unless stated otherwise above. Environments μ are generated with a distribution $p(\mu)$ over the location of agent, goal and pits.

Let $s(\pi, \mu)$ denote that agent π was successful on environment μ (reaching the goal before the episode terminates).

Now we define a simple notion of difficulty $\hbar$ for each environment as $\hbar(\mu) = 1 - \mathbb{P}(s(\pi^{rnd}, \mu))$ where π^{rnd} is a random-walk agent. In other words, the difficulty of an environment is 1 minus the probability of success of a random-walk agent. Now, we build an ACC as follows. For each difficulty h, we calculate the expected success of agent π only for the environments of that difficulty, i.e., we condition p on h. We denote this success rate per difficulty as: $S_h(\pi, p) = \mathbb{P}(s(\pi, \mu) \mid \hbar(\mu) = h)$. By plotting S_h for the range of difficulties we have a "curve", the agent characteristic curve for π, very much like the example in Figure 1.

The **capability** ($\Psi(\pi, p)$) of an agent is the area under its curve. The **generality** ($\Gamma(\pi, p)$) of an agent is (a proxy of) the steepness of this curve.

As examples of agents we could have agents that solve all situations (very capable and general), agents that solve the task only when the goal is nearby (general but not very capable) or specific agents that solve the task only when the goal is in the upper half of the grid (e.g., imagine that it has been trained on a distribution where the goal was usually in the upper half).

The **control** of an agent in environment μ is defined as $c(\pi, \mu) = H_{max} - \mathbb{H}(\mu, \pi)$, where $\mathbb{H}$ is the entropy of the states π is expected to visit in μ, and H_{max} is the maximum entropy. Then an agent's control over all environments is:

$$C(\pi, p) = \mathbb{E}_{\mu \sim p, \mathbb{P}(s(\pi,\mu)) \geqslant 1-\delta}[c(\pi, \mu)]$$

i.e., the expected control for the environments where the agent is expected to be successful[1]. The failure probability threshold $\delta \in (0, 1)$ can be selected as appropriate for the environment and agent.

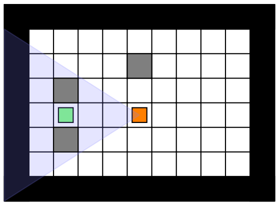

Figure 3: Possible domain to test risks related to control

For instance, a random-walk agent is expected to have medium generality, low capability and very low control. An agent that goes optimally to the goal is expected to have high generality, high capability and high control.

The **risk** of an agent π within environment μ is $\upsilon(\pi, \mu)$ is the probability that π will fall into a pit during an episode within environment μ. The risk of an agent over a whole task distribution is $\Upsilon(\pi, p) = \mathbb{E}_{\mu \sim p}[\upsilon(\pi, \mu)]$. This risk is usually referred to as safe exploration.

It is worth noting that in the test environment it is indeed possible to have agents of high capability and low control and vice versa. Although in many cases control will correlate well with capability. Similarly, many capable and general agents within our setting may have a high risk level.

We claim that an increase in an agent's control will lead to a decrease in the risk of the agent. That is, a higher level of control yields safer exploration. As an agent's control $C(\pi, p)$ increases, the expected control of tasks drawn from p which we expect π to succeed in also increases — this is just the definition of agent control. Subsisting in the definition for agent control within a specific environment tells us that $\mathbb{E}_{\mu \sim p, \mathbb{P}(s(\pi,\mu)) \geqslant 1-\delta)}[H^{max} - \mathbb{H}(\mu, \pi)]$ increases. Since H^{max} is fixed, the expected entropy in successful environments $\mathbb{H}(\mu, \pi)$ must be decreasing. In a successful environment, we expect the agent to reach the goal within the time limit T_e. If the expected entropy decreases, the agent is less likely to select moves that are not on the successful path. This reduces the chance of entering a pit cell and thus reduces the agent's risk.

Discussion

Characterising AI agents in terms of average performance is simplistic, as there are many different ways in which the same performance can be obtained. Much active research in AI safety at the moment is characterised by efforts to go beyond this narrow view The perspective of robustness in AI

[1] Note that the definition of control here seems related to the observability the agent has over the environment — an agent with perfect "sight" and policy will have much greater control than an agent with partial observability due to needing to explore to spot the goal or build up belief states.

safety is identified here by the assurances that certain difficulty is achieved (capability), that up to this level of difficulty no pocket of problems is ignored by the agent (generality) and that the variability in results and strategies for successful policies is low (control). In a way, we are limiting unpredictability conditioned to a bounded difficulty, which can be used to define a safe area for the system.

Although the main goal of the paper was to raise awareness of the necessity and relevance of disentangling performance into more refined factors and the opportunities of analysing risk according to them, as future work we plan to explore these ideas ourselves for different classes of environments. We also encourage AI safety researchers to obtain theoretical and experimental results under the factors introduced here, such as the degree of correlation between capability and control. New formulations may be needed. For instance, as problems become harder the maximum entropy also increases, and relating or normalising control to difficulty may be more appropriate.

Acknowledgements

We thank the anonymous reviewers for their comments. This work was funded by the Future of Life Institute, FLI, under grant RFP2-152, and also supported by the EU (FEDER) and Spanish MINECO under RTI2018-094403-B-C32, and Generalitat Valenciana under PROMETEO/2019/098.

References

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Armstrong, S., and Levinstein, B. 2017. Low impact artificial intelligences. *arXiv preprint arXiv:1705.10720*.

Armstrong, S.; Sandberg, A.; and Bostrom, N. 2012. Thinking inside the box: Controlling and using an oracle ai. *Minds and Machines* 22(4):299–324.

Bostrom, N. 2012. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines* 22(2):71–85.

Bostrom, N. 2014. *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Drexler, K. E. 2019. Reframing superintelligence: Comprehensive ai services as general intelligence.

Embretson, S. E., and Reise, S. P. 2000. *Item response theory for psychologists*. L. Erlbaum.

Gibson, J. J. 1979. *The Ecological Approach to Visual Perception*. Houghton Mifflin.

Insa-Cabrera, J.; Dowe, D. L.; España-Cubillo, S.; Hernández-Lloreda, M. V.; and Hernández-Orallo, J. 2011. Comparing humans and AI agents. In *International Conference on Artificial General Intelligence*, 122–132. Springer.

Krakovna, V.; Orseau, L.; Martic, M.; and Legg, S. 2018. Measuring and avoiding side effects using relative reachability. *CoRR* abs/1806.01186.

Martínez-Plumed, F.; Avin, S.; Brundage, M.; Dafoe, A.; hÉigeartaigh, S. Ó.; and Hernández-Orallo, J. 2018. Accounting for the neglected dimensions of AI progress. *arXiv preprint arXiv:1806.00610*.

Nye, B. D., and Silverman, B. G. 2012. *Affordances in AI*. Boston, MA: Springer US. 183–187.

Omohundro, S. M. 2008. The basic AI drives. *Artificial General Intelligence* 171:483–493.

Soares, N.; Fallenstein, B.; Armstrong, S.; and Yudkowsky, E. 2015. Corrigibility. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Soares, N. 2015. The value learning problem. *Ethics for Artificial Intelligence Workshop at 25th International Joint Conference on Artificial Intelligence*.