Paradigms of Artificial General Intelligence and Their Associated Risks

Abridged white paper/proposal for information only
Submitted 22 May 2018
Uploaded July 2019

Authors: Jose Hernandez-Orallo, Sean O hEigeartiagh
Centre for the Study of Existential Risk
University of Cambridge, UK

1. MOTIVATION AND STATE OF THE ART

The first Asilomar principle states: "the goal of AI research should be to create not undirected intelligence, but beneficial intelligence". Artificial *general* intelligence (AGI), however, is specifically characterised by its *generality*, which would allow AGI systems to be applied to a wide range of applications and display a variety of behaviours. Consequently, it seems that the greater the generality of a system, the more difficult it will be to *direct* its behaviour. Unfortunately, we lack proper tools to understand the relation between the generality of a system, its capacity, and the resources it uses, to the question of how these factors affect its control, potential and predictability and, accordingly, several safety issues.

To develop these tools, we must contemplate the **forms that AGI could take under different paradigms** in the first place. AI agents can be designed, trained and conditioned in many different ways. The design includes the architecture, principles and knowledge the system has initially, the training involves the way the system learns or is taught, with the information provided, and the conditioning involves the way feedback or motivation is used to steer or align the behaviour of the system.

One simple paradigm is an oracle AI. Despite originating in the early days of computer science, it is still the common model for supervised models, which underlie many decision-making systems in a myriad of applications. A system only taking questions and producing answers could make mistakes, be unreliable or replicate an unfair bias. But, on top of this, it has been hypothesised that a powerful future oracle might potentially be very harmful by manipulating humans, interpreting questions in unintended ways, or trying to escape constraints placed on it (Armstrong et al. 2012, Armstrong 2017).

The most common paradigm for autonomous systems nowadays is reinforcement learning (Sutton and Barto 1998), for which numerous issues and solutions have been analysed; these include corrupted rewards, interruptibility, and side effects, among others. This paradigm has many variants, some of them motivated by research in AI safety, such as inverse reinforcement learning (Ng and Russell 2000), learning from preferences (Christiano et al. 2017) or the recent inverse reward design paradigm (Hadfield-Menell et al. 2017).

There are settings with very different kinds of feedback to condition an AI system, such as adversarial settings, in the form of GANs, or "Turing learning" (Groß et al. 2017). Instead of correct answers or rewards, a system can be given the right behaviours, such as in learning by demonstration or imitation learning (Schaal 1999, Ho and Ermon 2016), and learning by human control or traces (Schaul et al. 2015). Other paradigms minimise or eliminate the reliance on external rewards or even feedback, such as knowledge seeking agents (Orseau et al. 2013) or intrinsic motivation (Bellemare et al. 2016).

Some of these paradigms have been used to develop top-down theories of what inductive inference or even artificial (general) intelligence should be. We have Solomonoff prediction in the first place, which is sequential (and hence under the oracle paradigm, Solomonoff 1964). But we also have interactive extensions (Hernandez-Orallo 2000b) or its adaptation to a reinforcement learning setting (Hutter 2005, Aslanides et al. 2017). In all of these cases, the agent is *separated* from the environment, and some safety issues may not appear compared to the case where the agent is part of the environment —a paradigm known as naturalised induction (Soares 2014, Soares and Fallenstein 2017). Some of these paradigms have

to be reanalysed to consider resource-bounded systems or AI systems that are subject to other real-world constraints, and to bridge the gap between the short and the long-term scenarios.

Metalearning, transfer learning, curriculum learning, incremental learning, machine teaching (Zhu 2015), etc, could also be considered to be potential key components of an AGI model, and could be combined with other paradigms, especially during the training stage. For instance, machine teaching can be combined with Solomonoff induction (Hernández-Orallo and Telle 2018), or planning with gradient descent (Srinivas et al. 2018). Also, the paradigms can be affected by the underlying technologies, with deep learning being a whole ecosystem of techniques. It may be that some of these current techniques could lead to powerful AGI systems without additional major breakthroughs and insights (Christiano 2016); it may also be that these techniques lead to powerful future systems that nonetheless lack some abilities required for them to be AGI under particular definitions.

There are additional issues to explore in multi-agent scenarios. If several agents share the same environment, even using the same paradigm, the emergent interactions may result in new safety issues or invalidate the results for the single agent case (Hadfield-Menell et al. 2016, Guerraou et al. 2017). Conversely, the distribution of multiple agents may make the system more robust (Christiano 2018). A particular case is AI on demand or cognition *as a service* (Spohrer et al. 2015), where a platform provides a wide range of services. In some cases, the particular agents are not allowed to learn or keep memory between services, but in other cases they can learn at the backstage in a more controlled way.

For each and all of these paradigms and combinations, many properties can be analysed, such as efficiency or convergence. In this project we are interested in the analysis of **the risks of AGI**, covering five previously identified **safety** areas —"side effects", "goal hacking", "oversight", "exploration" and "distributional shift" (Amodei et al. 2016)—, but also including the **malicious use** of AI (Brundage & Avin et al. 2018), where the goal is misuse in the first place. We aim to cover the whole landscape of AI Safety (Mallah 2017, Everitt et al. 2018).

However, from the previous literature review, it is clear that the paradigms are becoming more complex and more hybrid. As a result, the theoretical results (and especially positive theoretical results) are rare; each flaw requires a patch, which in turn creates new issues, in aeternum (Hibbard 2014). Also, if we consider Turing-complete systems, we face Rice's Theorem: no interesting property, including safety statements, can be determined for all possible situations. Partly because of this, empirical research has begun to take a more prominent role, with the introduction of AI Safety Benchmarks (Leike et al. 2017) mimicking the proliferation of benchmarks and the relevance of evaluation in AI (Castelvecchi 2016, Hernández-Orallo 2017a, 2017b, Hernández-Orallo et al. 2017). Either theoretically or empirically, instead of analysing each issue for each particular paradigm, **we are interested in some criteria that allow us to look at the landscape in a more comprehensive, gradual and accessible way.**

There are not many encompassing criteria in this direction. Three remarkable exceptions exist. The first one is the general view of **value alignment** (Russell et al. 2015), which is prominent in the agenda of AI safety, and triggers the discussion of the trade-off between pristine ethical principles and a more pragmatic view of what the good values are, relative to a social or historical moment. The second overarching criterion is a measurement of **impact**, quantifying those agents with low or high impact (Armstrong & Levinstein 2017). Again, we see the emergence of trade-offs, as low-impact AI may limit the magnitude of beneficial impacts. The third criterion is **autonomy**, which is usually understood in terms of independence, or taken for granted, rather than seen gradually according to the levels of supervision for a task that is required for an AI system to perform safely (Alexander et al. 2018).

There are two other criteria that have not been applied systematically to the issues of AI safety and the risks of AGI in particular: they are the **capability** of the system and its **generality**. It is true that for some paradigms the issues can appear independently of their capability and generality, because of the nature of a framework. For instance, even a simple reinforcement learning agent can potentially end up resisting being disconnected (Soares et al. 2015, Orseau and Armstrong 2016), but these situations may be more or less likely depending on these two dimensions. As we will argue below, these two criteria are probably the most relevant dimensions for the analysis of intelligence, and hence AGI. However, despite their importance, we lack good metrics of capability, and the situation is even worse for generality. It is even unclear whether generality is defined by a diversity of capabilities and tasks, or whether it is defined by a diversity of goals. Would hypotheses like 'basic AI drives' (Omohundro 2008) and the orthogonality thesis (Bostrom 2012) hold for systems with different levels of capability and generality, and developed under

different research paradigms? With proper metrics and understanding, we could reframe the agenda of AGI safety as a function of both capability and generality.

In summary, for all these paradigms, once several safety issues are categorised as possible**, we need a more fine-grained assessment of their plausibility, in terms of generality** (range of tasks and abilities covered by the systems) **and capability** (the performance profile for each of them). In other words, as AI systems become more general and more capable, it is reasonable to think that safety problems will be amplified and new challenges will emerge, as a function of these two dimensions.

Let us first look at **capability**. From an anthropocentric point of view, if the capability of a system starts resembling that of a human, we may expect all the dangerous situations humans can create, either by accident or maliciously. Second, a specific problem of AGI is the possibility of superintelligence (Bostrom 2017), especially if it creates a huge gap beyond human capabilities, or may even lead to phenomena such as an intelligent explosion (Good 1966), the singularity (Kurzweil 2010) and related forecasts (Armstrong et al. 2014).

However, there is still very **poor understanding of how to measure the capability of a system and, especially, how to link this capability, or maximum performance, to the required resources**. This can be done at a very theoretical level (Sotala, K. 2017, Turchin 2018), even linking it to physics and energy (Hernandez-Orallo 2018). But it can be seen more pragmatically as the ratio between performance of particular techniques such as deep learning, and the cost or consumption of their implementation on several computer architectures, such as FPGA, ASIC, CPU or GPU (Reagen et al. 2017, Hwang 2018, Amodei and Hernandez 2018). Other more neglected resources (Martínez-Plumed et al. 2018b) can be identified when the ultimate goal is performance for a single task. In these cases, the representativeness of the problem, as well as the replicability of results (or the cost to apply to other situations) is compromised because of an overly strong focus on task performance. This is also related to Goodhart's law and some of the safety problems an overemphasis on a metric can cause (Manheim and Garrabrant 2018).

In the context of AGI, analysis has also been done of "capability amplification", suggesting that in some circumstances increased capability may represent a solution rather than a problem from a safety perspective. Under some conditions, "*a policy that 'wants' to be aligned, in some sense —allowing it to think longer— may make it both smarter and more aligned*" (Christiano 2017). Indeed, in modern-day systems a lack of capability is rarely seen as a potential risk, but the use of insufficiently capable systems for complex situations may lead to problems. Incompetent systems are dangerous. All this suggests a possible **non-monotonic relation between capability and risks**, which warrants more systematic investigation.

Capability alone, as a monolithic concept, may fall short in providing a characterisation or modulation of risks. The ***generality*** of the system and how adaptable it is for a range of situations is not only intrinsic to learning (and hence intelligence) but has been a dimension of analysis from the early days of AI to the present day (Martínez-Plumed et al. 2018a). This draws upon similar previous debates around the concept of general intelligence in humans (Spearman 1927, Jensen 1998, Detterman 2002) and other animals (Burkart et al. 2017). Generality was already a pivotal issue for the early *general problem solver* (Newell 1959), a research programme that was abandoned for other narrower, but more successful, systems. Nonetheless, general intelligence returned as a central concept for the field by the time of McCarthy's Turing Award speech on "generality in AI" (McCarthy 1989) and more recently under the term Artificial General Intelligence (Adams et al. 2012).

AGI is frequently understood as Human-Level Machine Intelligence (HLMI) or Human-Level AI (HLAI), (e.g., Besold and Schmid 2016), defined as 'AI as good or better than humans at all cognitive tasks'. However, this definition does not clarify what human-level intelligence is, what level of generality it shows and how tasks or humans are selected. Variants such as High-Level Machine Intelligence (keeping the acronym, HLMI) still include "human" in their definition (Müller and Bostrom 2016, Grace et al. 2017). Alternatively, a less anthropocentric perspective of AGI that we pursue in this project may deviate from HLAI in several ways, such as being very general but less capable than humans, or covering some tasks that humans cannot cover.

The relevance of generality for *safety* in AI has been recognised more recently. For instance, Krakovna (2018) puts it this way: "*For many narrow systems and narrow applications where you can sort of foresee all the ways in which things can go wrong, and just penalize all those ways or build a reward function that avoids all of those failure modes, then there isn't so much need to find a general solution to these problems.*

While as we get closer to general intelligence, there will be more need for more principled and more general approaches to these problems".

Limiting generality (or trading it against capability or resources) seems currently unattainable as a governance tool. Firstly, there is a strong economic pressure towards more general systems as they may eventually become cheaper and more broadly applicable than specialised systems requiring lengthy development, especially in the context of automations and jobs (Frey and Osborne 2017, Nedelkoska and Quintini 2018, Fernández et al. 2018). Secondly, it is unclear where and how to put a limit on generality.

Generality as a concept is difficult to define. For instance, under the oracle or cognitive services paradigms, we may consider a *general* personal assistant. But are we referring to a wide repertoire of pre-defined tasks and services, or do we mean a system that can be trained to do any task? Can subgoals be general even with a very restrictive main goal? And, in a social context, can we understand generality as capturing the beliefs, desires and intentions (or values) of a wide range of other users (minds), or just those that are similar to the modelling agent? (Rabinowitz 2018).

A further problem is that **generality is usually conflated with the very concept of intelligence or capability**. For instance, AGI, as a term, is sometimes used as synonym for human-level machine intelligence or even superintelligence. This conflation has been reinforced by notions of intelligence as good performance on *all* cognitive tasks, or all tasks humans can do, leading to different interpretations. For instance, universal intelligence (Legg and Hutter 2007) is a very elegant formulation of a notion of intelligence covering all computable tasks in an interactive setting, and as an alternative to the assumptions of the no-free-lunch theorems (Wolpert & MacReady 1997). The formal definition is usually (mis-)interpreted as performance in a wide range of environments, despite its subjectivity (Hibbard 2009, Hernández-Orallo and Dowe 2010, Leike & Hutter 2015, Hernandez-Orallo 2017a). Indeed, given limited resources (compute, time), a system will only be able to cover a small subset of all possible tasks. As a result, even under these frameworks, any system would be specialised. How can we think of a metric of generality, different from capability, that is consistent and meaningful for resource-bounded agents?

2. HYPOTHESES AND GOALS

One important hypothesis of this project is the **possibility of decoupling generality from capability for all the AGI paradigms**. For instance, if we look at the variance of results —more variance, less generality— and aggregated scores of a set of systems on a range of tasks, we see quite a constrained picture (Figure 1, left), where variance and score are strongly correlated. In the end, the reciprocal of variance may not be a good measure of generality. However, there is an important source of evidence where generality can be seen in a different way and can be decoupled from capability. This is item response theory, IRT (Embretson and Reise 2000), which is also being applied to machine learning and AI (Martínez-Plumed et al. 2016). IRT extracts latent factors that are able to explain the behaviour of tasks (items) and agents (persons). Person characteristic curves show accomplishment vs. difficulty, in the form of sigmoid curves, where the location could be seen as the capability and the slope of the curve could be seen as the generality (higher slope, higher generality, see Figure 1, right).

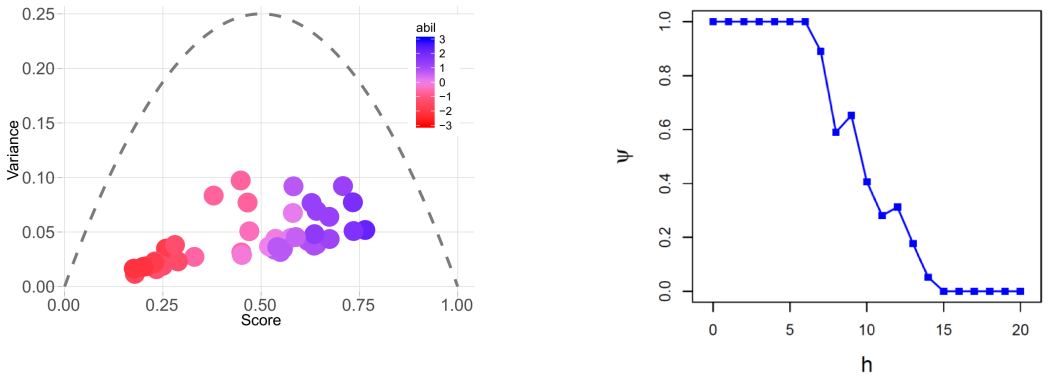

Figure 1. Left: Variance vs. average normalised score for twenty AI techniques studied for the ALE benchmark (Atari games). The dashed grey curve is the variance of a Bernoulli distribution (the worst case). Right: an agent characteristic curve, with resources (hardness or difficulty) shown on the x-axis and accomplishment on the y-axis. Data from (Hernandez-Orallo 2000).

Looking at Figure 1 (right) we see that covering all tasks is impossible for a system with limited resources: very difficult tasks are unattainable. Consequently, it seems more efficient if a general system focuses on all the easy problems first, before using resources or architecture for the more complex ones. This is especially the case if the system is going to be deployed on situations where we do not know the a priori distributions of the tasks. In other words, we can see generality as covering a wide range of tasks *up to a given level of difficulty or resources*.

Another hypothesis is that **both generality and capability can be used to find trade-offs, by analysing the enhancement of capability and generality as a function of required resources**. We have to consider that, in the first place, decoupling generality and capability does not mean that they may be independent for *particular* AGI systems. A pressure to achieve simple tasks could lead to some correlations with capability. For instance, as we can see in Figure 2, a cognitive enhancement for a system increases its capability more efficiently if it focuses on easy tasks.

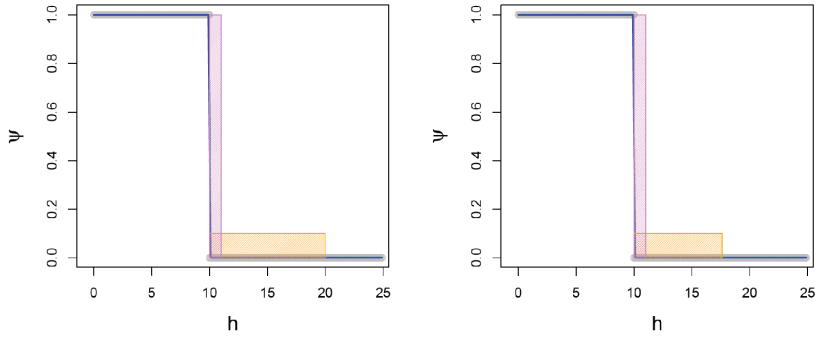

Figure 2. Two agent characteristic curves (with resources on the x-axis and accomplishment on the y-axis) with a domain-general (vertical violet rectangle) or a domain-specific (horizontal orange rectangle) cognitive enhancement. Left: both rectangles cover the same area (and capability). Right: both rectangles imply the same effort in resources, but the specific one (orange rectangle) now has a smaller area, and hence less increase in capability than the general one (violet rectangle).

What these plots help show more explicitly is the relation between capability, generality and resources. This can be exploited for safety issues, especially for the analysis of self-improvement, superintelligence and growth, in terms of different curve shapes.

Finally, a third hypothesis is that **using resources (or difficulty) as a base for deriving capability and generality metrics leads to a meaningful notion of intelligence that can be applicable to all paradigms**, especially in experimental settings when considering a range of tasks. This is illustrated in Figure 3 (bottom), where we can consider *all* possible tasks with a uniform distribution of difficulties, leading to a way of considering all possible tasks that does not fall into many of the problems of other formalisms, including the no-free-lunch theorems.

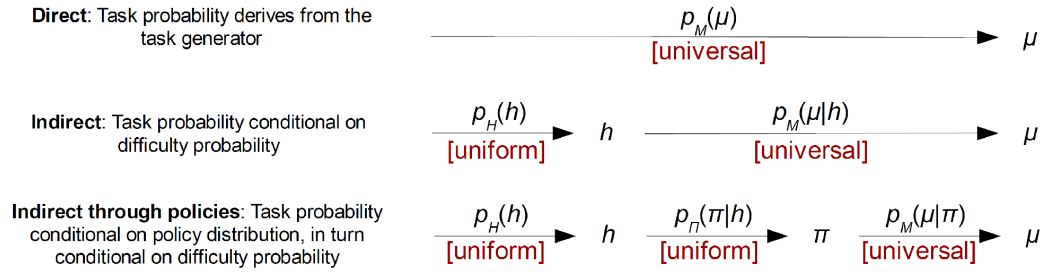

Figure 3. Top: the probability of a task (μ) derives (inversely) from the complexity of its description. If a universal distribution is used for this, this becomes Legg and Hutter's universal intelligence (2017). Note that the description of a task does not say much about the difficulty (h) of the solution policy (π), especially in interactive environments, Middle: we start with a distribution of difficulties (h), and tasks are conditional to that difficulty (Hernandez-Orallo 2000). Bottom: starting with difficulties then we derive policies (π) of that difficulty. Finally, tasks are conditional to the policy (Hernandez-Orallo 2017a).

The explicit connection with resources (*h*) as a starting point suggests that if risks can be quantified and controlled by a proper modulation of resources for capability and generality, we could derive new tools for AI governance, by exploring ways of tracking computing power and avoiding concentrations.

Given these considerations, we formulate the following main goals of the project:

1. **Catalogue the paradigms of AGI**, according to the principles the system is based on, the way its behaviour is conditioned and how learning takes place, and map the risks (safety issues and malicious use) over these paradigms.
2. **Derive metrics of generality and capability** that can be used comprehensively and effectively over multiple paradigms, and link these measures with resources, in the context of capability amplification, self-improvement and intelligence growth.
3. **Identify more detailed desirable intelligence profiles,** and connect with the generality and capability metrics, in the context of disruptions related to automation and oversight (**replacement risks**), as well as mind modelling and cognitive diversity (**domination risks**).
4. **Propose solutions** for some of these paradigms, **according to trade-offs and limitations** (of generality, capability or resources, autonomy, differences or diversity in mind-modelling capacity) that minimise the risks.

The project will deliver white papers and survey papers for the use by the AI safety community, but will also promote new research questions through technical papers and workshops at major conferences. Dissemination to the research community, policymakers and the public will be leveraged by CSER and CFI's excellent global outreach networks.

Work packages:

WP1: Management. Goals: Overall coordination of the project, its internal synergy, its monitoring and accountability, human resources and formal arrangements, reporting, risk assessment and validation.

WP2: Outreach. Goals: Manage the external relations of the project, its dissemination and publicity, the engagement of the community and the creation of public events and organisations.

WP3: Generality. Goals: Derive conceptions and measures of generality and capability that can be used for the evaluation of AI systems, and are compatible with natural systems.

WP4: Paradigms and Risks. Goals: Produce a taxonomy of paradigms in AGI, and a comprehensive mapping from these paradigms to the risks of AI, following FLI's AI safety landscape (Mallah 2017) and other surveys.

WP5: Resources and growth. Goals: Create a list of criteria in which resources are linked to the growth of generality and capability, in order to reframe intelligence growth, augmentation and superintelligence accordingly.

WP6: Safe landmarks. Goals: Locate those combinations of paradigms, levels and trade-offs of generality and capability that are safer, individually and collectively.

SP7: Replacement risks. Goals: Determine when a system A can replace a system B, according to the capability and generality of both systems, and what risks this may imply, especially in the workplace.

SP8: Domination risks. Goals: Determine when a system A can dominate a system B, according to their capability and generality, and the implied risks, especially in social environments and in terms of agent diversity.

3. VISION AND LONG-TERM-IMPACTS

This project presents a **new perspective on the potential risks of artificial general intelligence in a more comprehensive, analytical and rigorous way**, at both the theoretical and empirical levels, by clearly delineating the paradigm of the agent, and the metrics of capability and generality. This will have a short-term impact on research agendas during and immediately after the end of the project.

Ultimately, the main vision is to analyse risks in terms of more abstract criteria, such as generality and capability, as an alternative or complementary conceptual framework to autonomy, alignment and impact, instead of endless combinations of particular paradigms and categories of risks.

In practical terms, the project will influence the assessment of the three main areas of risks: safety (how and when the system may go wrong), misuse (security; and how the system may be used for malicious purposes) and competition (do resources linked to growth in capability and generality equalise or

destabilise competition between actors). **We aim to break the analytical schism between short-term and long-term risks**, and see a gradation in the paradigms from specialised AI to AGI.

As long-term impacts, we enumerate the following:

- **A more comprehensive, and non-anthropocentric, characterisation of the forms that AGI could take**, and how to conceptualise it according to new metrics of capability and generality. This is *foundational* and will thus have implications on all AGI research, especially around safety issues.
- A taxonomy of paradigms in research aimed at achieving AGI, covering many different paths towards general intelligence, which can bridge the communities in AI and safety, and render a **better way of structuring and connecting research in AGI safety, as more capable and general systems are developed**.
- A series of benchmarks, metrics and tools, including tests and software, for AI safety research, which will **allow researchers to include the metrics of generality and capability, versus resources, in their algorithms and optimise accordingly**.
- A different, less monolithic, view of augmentation, self-improvement and superintelligence, in height and breadth, which can **move the debate forward from simplistic views of exponential explosion and the singularity** (e.g., European Parliament, Bentley 2018).
- A different view on the future of work, in a the context of replacement risks. Instead of probabilities per profession, we will analyse how cognitive profiles affect the whole labour market and what **professional profiles are ranked with highest risks *if automated***. (spin-off contribution)
- A grounded connection between domination and risks, **clarifying when a system can dominate another system or a group of diverse individuals**, and its connections with social intelligence, manipulation and diversity (spin-off contribution).

In order to maintain the generation of research, career development and exploitation of the human assets and synergies of the project once finished, we have devised two stretch packages to be converted into spin-off projects. We also plan more ambitious projects (e.g., ERC projects in Europe) and the consolidation of the workshops as a periodic event. AGI safety is extremely clustered, mostly concentrated in the USA (as seen in Baum 2017, Fig. ESI). This project will aim to increase the diversity (in terms of people and approaches) of the ecosystem of AGI research and associated risks, especially through the connections of CSER and CFI, and the European Commission (HumaInt).

Overall, this project will contribute new concepts to AGI and safety, and more powerful tools for the AI research and governance communities, which will be technically rigorous yet intelligible and beneficial for society.

4. REFERENCES

- Adams, S., Arel, I., Bach, J., Coop, R., Furlan, R., Goertzel, B., Hall, J.S., Samsonovich, A., Scheutz, M., Schlesinger, M., Shapiro, S.C., and Sowa, J. (2012) Mapping the landscape of human-level artificial general intelligence. AI magazine, 33(1):25–42.
- Alexander, R. D., Ashmore, R., and Banks, A. (2018) "The State of Solutions for Autonomous Systems Safety", Safety-Critical Systems Symposium.
- Amodei, D., Olah, C., Steinhardt, J. Christiano, P., Schulman, J. and Mané, D. (2016) "Concrete problems in AI safety." *arXiv preprint arXiv:1606.06565*.
- Amodei, D., Hernandez, D. (2018) "AI and compute" https://blog.openai.com/ai-and-compute/,
- Armstrong, S. (2017). Good and safe uses of AI Oracles. *arXiv preprint arXiv:1711.05541*.
- Armstrong, S., & Levinstein, B. (2017). Low Impact Artificial Intelligences. *arXiv:1705.10720*.
- Armstrong, S., Sotala, K., & Ó hÉigeartaigh, S. S. (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.
- Armstrong, Stuart, Anders Sandberg, and Nick Bostrom (2012) "Thinking inside the box: Controlling and using an Oracle AI" *Minds and Machines* 22.4: 299-324.
- Aslanides, J., Leike, J., & Hutter, M. (2017). Universal reinforcement learning algorithms: Survey and experiments. arXiv preprint arXiv:1705.10557.
- Baum, S. D. (2017) A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy. Tech. rep. November. Global Catastrophic Risk Institute, pp. 1-99.

- Bellemare, M., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., & Munos, R. (2016). Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems* (pp. 1471-1479).
- Bentley, P. J. (2018) "The Three Laws of Artificial Intelligence: Dispelling Common Myths", in European Parliament Think Tank report "Should we fear artificial intelligence?".
- Besold, T. R., and Schmid, U. (2016) "Why Generality Is Key to Human-Level Artificial Intelligence." Advances in Cognitive Systems, (4): 13-24.
- Bhatnagar S, Alexandrova A, Avin S, Cave S, Cheke L, Crosby M, Feyereisl J, Halina M, Loe BS, Ó HÉigeartaigh S, Martínez-Plumed F, Price H, Shevlin, H, Weller A, Winfield A, Hernández-Orallo J (2018) "Mapping Intelligence: Requirements and Possibilities", in V. Müller, (ed) Proceeding sof the Philosophy and Theory of Artificial Intelligence, Leeds.
- Bieger, J., Thórisson, K. R. and Wang, P. (2015) "Safe Baby AGI." In International Conference on Artificial General Intelligence, pp. 46-49. Springer.
- Bostrom, N. (2012). The Superintelligent Will: Motivation and Instrumental Rationality in Advance Artificial Agents. Minds and Machines, 22(2), 71-85.
- Bostrom, N. (2014). Superintelligence. Dunod.
- Brundage, M. & Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A, Scharre, P., Zeitzoof, T., Filar, B., Anderson, H., Roff, H., Allen, G.C., Steinhardt, J. Flynn, C., Ó hÉigeartaigh, S., Beard, S., Belfield, H., Faruhar, S., Lyle, C., Crootof, R., Ewans, O., Page, M., Bryson, J., Yamposkiy, R., Amodei, D. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. *arXiv:1802.07228*.
- Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. Science, 358(6370), 1530-1534.
- Burkart, J.M., Schubiger, M.N., and van Schaik, C. P. (2017) The evolution of general intelligence, Behavioral and Brain Sciences, 40.
- Burton, E., Goldsmith, J., Koenig, S., Kuipers, B., Mattei, N., & Walsh, T. (2017). Ethical considerations in artificial intelligence courses. arXiv preprint arXiv:1701.07769.
- Castelvecchi, D. (2016) "Tech giants open virtual worlds to bevy of AI programs", Nature, 14 Dec.
- Christiano, P. (2016) "Prosaic AI alignment".
- Christiano, P. (2017) "AlphaGo Zero and capability amplification"
- Christiano, P. (2018) "Universality and security amplification",
- Christiano, P., Leike, J., Brown, T., Martic, M., Legg, S. and Amodei, D. (2017) "Deep reinforcement learning from human preferences." In *Advances in Neural Information Processing Systems*, pp. 4302-4310. 2017.
- Detterman, D.K.. (2002) General intelligence: Cognitive and biological explanations. In R. J. Sternberg and E. L. Grigorenko, The general factor of intelligence: How general is it? pp 223–243.
- Embretson, S.E. and Reise, S. P. (2000) Item response theory for psychologists, L. Erlbaum.
- Everitt, T., Lea, G. Lea, and Hutter, M.. (2018)"AGI Safety Literature Review". IJCAI.
- Fernández-Macías, E., Gómez, E., Hernández-Orallo, J., Loe, B.S., Martens, B., Martínez-Plumed, F., Tolan, S. (2018) "A multidisciplinary task-based perspective for evaluating the impact of AI autonomy and generality on the future of work", submitted to AEGAP@IJCAI2018.
- Frey, C.B., Osborne, M. (2017) The future of employment: how susceptible are jobs to computerisation? Technological Forecasting and Social Change, 114:254–280.
- Good, I. J. (1966) Speculations concerning the first ultraintelligent machine In Advances in computers (Vol. 6, pp. 31-88). Elsevier.
- Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2017) When will AI exceed human performance? Evidence from AI experts. arXiv:1705.08807.
- Groß, R., Gu, Y., Li, W., & Gauci, M. (2017). Generalizing GANs: A Turing Perspective. In *Advances in Neural Information Processing Systems* (pp. 6319-6329).
- Guerraoui, R., Hendrikx, H., & Maurer, A. (2017). Dynamic Safe Interruptibility for Decentralized Multi-Agent Reinforcement Learning. In Advances in Neural Information Processing Systems (pp. 129-139).
- Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. In Advances in neural information processing systems (pp. 3909-3917).
- Hadfield-Menell, D., Milli S., Abbeel, P., Russell, S.J. and Dragan, A. (2017) "Inverse reward design." In *NIPS*, pp. 6768-6777.
- Hernández-Orallo, J. (2000a) Beyond the Turing Test. J. Logic, Lang. & Information, 9(4):447–466.
- Hernández-Orallo, J. (2000b). On the computational measurement of intelligence factors, PerMIS, 72-79, NIST.
- Hernández-Orallo, J. (2015) "C-tests revisited: Back and forth with complexity". In J. Bieger, B. Goertzel, and A. Potapov, editors, Artificial General Intelligence - AGI 2015, Proceedings, pages 272–282. Springer.
- Hernández-Orallo, J. (2017a) *The measure of all minds: evaluating natural and artificial intelligence*. Cambridge University Press, 2017.
- Hernández-Orallo, J. (2017b) "Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement." Artificial Intelligence Review, October 2017b, Volume 48, Issue 3, pp 397–447.
- Hernández-Orallo, J. (2018) "Intelligence without Measure" *Nature Physics*, to appear.

- Hernández-Orallo, J., and Dowe, D.L. (2010) "Measuring universal intelligence: Towards an anytime intelligence test." *Artificial Intelligence* 174.18: 1508-1539, 2010.
- Hernández-Orallo, J., Baroni, M.; Bieger, J.; Chmait, N.; Dowe, D.L.; Hofmann, K.; Martínez-Plumed, F.; Strannegård, C.; Thórisson, K.R. (2017) "A New AI Evaluation Cosmos: Ready to Play the Game?", AI Mag..
- Hernández-Orallo, J., and Telle, J.A.. (2018) "Finite Biased Teaching with Infinite Concept Classes" *arXiv:1804.07121* (2018).
- Hibbard, B. (2009). Bias and no free lunch in formal measures of intelligence. Journal of AGI, 1(1), 54.
- Hibbard, B. (2014). Ethical artificial intelligence. *arXiv preprint arXiv:1411.1373*.
- Ho, J., & Ermon, S. (2016). Generative adversarial imitation learning. In Advances in Neural Information Processing Systems (pp. 4565-4573).
- Hutter, M. (2005). *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Springer.
- Hwang, T., (2018) "Computational Power and the Social Impact of Artificial Intelligence".
- Jensen, R. (1998) The g factor: The science of mental ability. Westport, Praeger..
- John McCarthy (1987) Generality in artificial intelligence. Communications of the ACM, 30(12):1030–1035.
- Krakovna, V. (2018) "Podcast: Navigating AI Safety – From Malicious Use to Accidents" https://futureoflife.org/2018/03/30/podcast-malicious-use-of-artificial-intelligence/
- Kurzweil, R. (2010). The singularity is near. Gerald Duckworth & Co.
- Legg, S., & Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. Minds and Machines, 17(4), 391-444.
- Leike, J., & Hutter, M. (2015). Bad universal priors and notions of optimality. COLT 1244-1259.
- Leike, J., Martic, M., Krakovna, V., Ortega, P. A., Everitt, T., Lefrancq, A., Orseau, L., Legg, S. (2017). AI Safety Gridworlds. *arXiv preprint arXiv:1711.09883*.
- Mallah, R. (2017) "The Landscape of AI Safety and Beneficence Research: Input for Brainstorming at Beneficial AI 2017".
- Manheim, D. and Garrabrant, S. "Categorizing Variants of Goodhart's Law" (2018) arXiv:1803.04585.
- Martínez-Plumed, F., Loe, B. S., Flach, P., Ó hÉigeartaigh, S., Vold, K.; Hernández-Orallo, J. (2018a) "The Facets of Artificial Intelligence: A Framework to Track the Evolution of AI", IJCAI.
- Martínez-Plumed, F., Avin, S., Brundage, M., Dafoe, A., Ó hÉigeartaigh, S., Hernández-Orallo, J. (2018b) Accounting for the Neglected Dimensions of AI Progress, submitted.
- Martínez-Plumed, F., Prudencio, R.B.C., Martínez Usó, A. and Hernández-Orallo, J. (2016). "Making sense of item response theory in machine learning." ECAI (Best paper award), 1140-1148.
- Mikolov, T., Joulin, A., & Baroni, M. (2016). A roadmap towards machine intelligence. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 29-61). Springer,
- Müller, V. C., and Bostrom, N. (2016) "Future progress in artificial intelligence: A survey of expert opinion." In Fundamental issues of artificial intelligence, pp. 555-572. Springer.
- Nedelkoska, L. and G. Quintini (2018), "Automation, skills use and training", OECD Social, Employment and Migration Working Papers, No. 202, OECD Publishing, Paris.
- Newell, A., Shaw, J.C. and Simon, H.A. (1959) Report on a general problem-solving program. IFIP, 256–264.
- Ng, A. Y., and Russell, S.J. (2000) "Algorithms for inverse reinforcement learning" *ICML* pp. 663-670.
- Omohundro, S. M. (2008). The Basic AI Drives. In P. Wang, B. Goertzel, & S. Franklin (Eds.), Artificial General Intelligence: Proceedings of the First AGI Conference (Vol. 171)
- Orseau, L., & Armstrong, M. S. (2016). Safely interruptible agents.
- Orseau, L., Lattimore, T., & Hutter, M. (2013). Universal knowledge-seeking agents for stochastic environments. In Intl. Conf on Algorithmic Learning Theory (pp. 158-172). Springer.
- Rabinowitz, N. C., Perbet, F., Song, H.F., Zhang, C., Eslami, S.M., and Botvinick, M. (2018) "Machine Theory of Mind." arXiv:1802.07740.
- Reagen, B., Adolf, R., Whatmough, P., Wei, G-Y and Brooks, D. (2017). Deep learning for computer architects. SL on Comp. Architecture, 12(4):1–123.
- Russell, S., Dewey, D., & Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. AI Magazine, 36(4), 105-114.
- Schaal, S. (1999). Is imitation learning the route to humanoid robots?. *Trends in cog. sciences*, *3*(6), 233-242
- Schaul, T., Quan, J., Antonoglou, I., & Silver, D. (2015). Prioritized experience replay. *arXiv:1511.05952*.
- Soares, N. (2014). *Formalizing two problems of realistic world-models*. Tech. Rep. MIRI.
- Soares, N., & Fallenstein, B. (2017). Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda. In *The Technological Singularity* (pp. 103-125). Springer.
- Soares, N., Fallenstein, B., Armstrong, S., & Yudkowsky, E. (2015). Corrigibility. In *Ws AAAI.*
- Solomonoff, R. J. (1964). A formal theory of inductive inference. Part I. *Information and control*, *7*(1), 1-22.
- Sotala, K. (2017). How feasible is the rapid development of artificial superintelligence?. Physica Scripta, 92(11).
- Spearman, C. (1927) The abilities of man: Their nature and measurement. Macmillan, New York.
- Spohrer, J., & Banavar, G. (2015) Cognition as a service: an industry perspective. *AI Magazine*, *36*(4), 71-86.

- Srinivas, A., Jabri, A., Abbeel, P., Levine, S., & Finn, C. (2018). Universal Planning Networks. arXiv:1804.00645.
- Sutton, Richard S., and Andrew G. Barto. (1998) *Introduction to reinforcement learning*. MIT press.
- Turchin, A. (2018) Levels of Self-Improvement in AI and their Implications for AI Safety.
- Wolpert, D. H., & Macready, W. G. (1997) No free lunch theorems for optimization. IEEE transactions on evolutionary computation, 1(1), 67-82,
- Zhu, X. (2015) Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In AAAI, pages 4083–4087.