Journal of Artificial General Intelligence 11(1) 1-4, 2020
DOI: 10.2478/jagi-2020-0001

Submitted 2019-10-26
Accepted 2019-11-07

Defining Artificial Intelligence: Resilient Experts, Fragile Geniuses, and the Potential of Deep Reinforcement Learning

Matthew Crosby M.CROSBY@IMPERIAL.AC.UK
Leverhulme Centre for the Future of Intelligence
Imperial College London
London, United Kingdom

Henry Shevlin HFS35@CAM.AC.UK
Leverhulme Centre for the Future of Intelligence
University of Cambridge
Cambridge, United Kingdom

Editors: Dagmar Monett, Colin W. P. Lewis, and Kristinn R. Thórisson

Abstract

Wang's definition of Artificial Intelligence is developed via careful and thorough abstractions from human intelligence. Motivated by the goal of building a definition that will be genuinely useful for AI researchers, Wang ultimately provides an agent-centric definition that focuses on systems operating with insufficient knowledge and resources. The definition captures many key components of intelligence, but we suggest that task success could play a slightly larger role. This brings the definition closer in line with our use of the term with animals and human experts, and also further aligns the definition's associated research framework with the subfield of deep reinforcement learning aimed at general intelligence.

1. Introduction

Wang (2019) proposes a working definition of artificial intelligence based on a system's ability to adapt to its environment under certain resource constraints (AIKR). The definition is accompanied by a useful exploration of different perspectives in AI that vary with respect to how they abstract from human intelligence. One *prima facie* challenge for this approach is the worry of anthropocentrism: the space of possible intelligences is vast, and anchoring the definition to humans risks blinding us to a large portion of it. However, to the credit of Wang's approach, humans are used as *exemplars* of the explicandum, not necessarily its sole arbiters.

We believe that Wang's definition picks out key elements of intelligence. The insufficiency assumption allows for a Principle-AI-based definition that does not fall into the trappings of extreme versions, such as AIXI (Hutter, 2005), that revolve almost exclusively around task solving. Its move away from purely capability-based definitions is also positive. However, we suggest that some reference to capabilities would still be of benefit to Wang's approach. We first suggest that this will bring it closer to our usage with animals and experts, then look at how an updated definition aligns with the intuitions of AI researchers working towards general intelligence in deep reinforcement learning.

2. Intelligence and insufficiency

We agree with Wang that the ability to deal efficiently with scenarios in which knowledge and resources are lacking is a key marker of intelligence. One fact that this definition captures elegantly is that intelligence is not just a matter of completing tasks: if one's goals are simple and the environment stable, it is possible to thrive via relatively simple strategies, or a 'Resilient Idiot' approach, as exemplified by organisms like nematode worms or sessile shellfish. However, the definition is arguably too restrictive as currently stated. In particular, we suggest it risks leaving out two types of intelligent system that we term 'Resilient Experts' and 'Fragile Geniuses'. We define a Resilient Expert as a system that has rich stores of knowledge and multiple redundant mechanisms for solving any problems it encounters. Much like the Resilient Idiot, the Resilient Expert simply does not encounter insufficiency or uncertainty. Unlike the case of the Resilient Idiot, however, this resilience is a hard-won achievement for the Resilient Expert, and is founded upon expensive investments in knowledge and resources.

A wide range of biological organisms that we are inclined to describe as intelligent might plausibly qualify as Resilient Experts. As a simple example, note the extremely robust navigational capacities of animals such as bees. Bees make use of environmental landmarks, track the location of the sun, calculate the polarity of light (useful on overcast days), and track the speed and vector of prior movement using dedicated neural assemblies (Gould and Gould, 1988; Stone et al., 2017). The resilience and complexity of the bee's complex navigational toolkit bespeaks a sophisticated and intelligent biological agent. This is in spite of the fact that the bee (at least qua navigation) rarely if ever has insufficient resources to carry out its tasks. We recognise that the case of the bee just provided involves a single task, namely navigation, and Wang notes that AIKR applies to "the overall situation, not on every task, as there are surely simple tasks for which the system's knowledge and resources are relatively sufficient." However, we think it reasonable to imagine that there could be Resilient Experts whose knowledge and resources were bountiful in every domain yet still qualified as intelligent.

A second form of intelligent system that Wang's definition of intelligence might not easily cover in its current form is the Fragile Genius. By a Fragile Genius, we mean a system that struggles with uncertainty and insufficiency, but which (intuitively) constitutes an instance of intelligence by virtue of specialising towards some particularly impressive or complex goal. Consider a brilliant but eccentric composer who writes symphonies of dazzling beauty, creativity, and complexity, but who is incapable of reliably feeding or clothing themselves or even obtaining materials for producing their compositions. They are wholly dependent on the cooperation of the external environment for their continued thriving and do not adapt well under AIKR conditions.

Most of us are Fragile Geniuses. Our way of life depends on rich cultural and technological knowledge and complex co-ordination and specialisation of roles. Without the scaffolding of our cultural knowledge and technology most of us would struggle to adapt to even basic tasks like obtaining food, constructing shelter, or treating injuries (Henrich, 2017). Nonetheless, it is surely false to suggest that the fragility of modern life is such that fewer demands are placed on our intelligence. Rather, the acquisition of rich cultural storehouses of knowledge and specialisation of individuals has enabled us to develop skills and proficiencies unthinkable for our neolithic ancestors, including such elevated achievements as quantum mechanics, aeronautical engineering, and the Baked Alaska.

While adaptation under AIKR is highly indicative of intelligence, some very intelligent agents—the Resilient Experts—have managed to avoid uncertainty and insufficiency all together via complex redundant systems, while others—the Fragile Geniuses—struggle in the face of these factors. Hence we would suggest that task complexity and task solving ability might be given some more prominent role in the definition, even if uncertainty and insufficiency remain the unifying theme.

3. Deep Reinforcement Learning and Artificial Intelligence

Measuring progress towards intelligence is hard, so AI research tends instead towards measurable tasks with determinate success conditions. This leaves current mainstream research strands at odds with the kind of definition Wang is proposing. However, with task capability and complexity even a small part of the picture, we believe that certain strands of current deep reinforcement learning research do qualify as working towards intelligence. The goal *is* learning that is lifelong, cumulative, open-ended, and multi-objective, it's just a long way away.

Wang suggests that the optimality of many machine learning algorithms goes against AIKR. Whilst many Deep Reinforcement Learning (DRL) algorithms are based on convergence proofs, Deep Learning usually involves non-linear approximations and DRL is often applied in situations where assumptions required for the proofs do not hold. It is commonly assumed that environments are fully observable Markov Decision Processes, but in practice this is rarely the case (Arulkumaran et al., 2017).

Lifelong and continual learning is a growing area of research in the DRL community, starting with methods to prevent catastrophic forgetting (Kirkpatrick et al., 2017), where neural networks will sometimes jump away from a favourable weight space and 'forget' everything they have previously learned. Whilst the environments used for research often do not deviate too far from standard machine learning (Lopez-Paz and Ranzato, 2017), progress is being made towards the introduction of new continual learning paradigms (Khetarpal et al., 2018). There are many open issues, and research is still in its infancy (Schaul et al., 2018), but first steps are being taken towards testing systems in finite, open, and real-time settings (Beyret et al., 2019). Many 'intelligent' AI researchers are working with a similar definition of artificial intelligence. It is perhaps ironic that their goal is frustrated by insufficient knowledge and resources.

4. Conclusion

Wang's definition picks out important components of intelligence and sets an interesting research agenda. It scores well on the criteria of similarity, exactness, fruitfulness, and simplicity, but could make stronger requirements on task capabilities. Doing so brings it closer to our usage for resilient experts, fragile geniuses, and even many DRL researchers.

Acknowledgments

This work was supported by the Leverhulme Centre for the Future of Intelligence, Leverhulme Trust, under Grant RC-2015-067.

References

Arulkumaran, K.; Deisenroth, M. P.; Brundage, M.; and Bharath, A. A. 2017. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine* 34(6):26–38.

Beyret, B.; Hernández-Orallo, J.; Cheke, L.; Halina, M.; Shanahan, M.; and Crosby, M. 2019. The Animal-AI Environment: Training and Testing Animal-Like Artificial Cognition. *arXiv preprint arXiv:1909.07483*.

Gould, J. L., and Gould, C. G. 1988. *The Honey Bee.* Scientific American Library.

Henrich, J. 2017. *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.

Hutter, M. 2005. *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Springer Science & Business Media.

Khetarpal, K.; Sodhani, S.; Chandar, S.; and Precup, D. 2018. Environments for Lifelong Reinforcement Learning. *arXiv preprint arXiv:1811.10732*.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences* 114(13):3521–3526.

Lopez-Paz, D., and Ranzato, M. 2017. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, 6467–6476.

Schaul, T.; van Hasselt, H.; Modayil, J.; White, M.; White, A.; Bacon, P.; Harb, J.; Mourad, S.; Bellemare, M.; and Precup, D. 2018. The Barbados 2018 List of Open Issues in Continual Learning. *arXiv preprint arXiv:1811.07004*.

Stone, T.; Webb, B.; Adden, A.; Weddig, N. B.; Honkanen, A.; Templin, R.; Wcislo, W.; Scimeca, L.; Warrant, E.; and Heinze, S. 2017. An anatomically constrained model for path integration in the bee brain. *Current Biology* 27(20):3069–3085.

Wang, P. 2019. On Defining Artificial Intelligence. *Journal of Artificial General Intelligence* 10(2):37.