Journal of Artificial General Intelligence 11(1) 1-3, 2020 Submitted 2019-10-01
DOI: 10.2478/jagi-2020-0001 Accepted 2019-11-07

Intelligence and Agency

Peter Lindes PLINDES@UMICH.EDU
University of Michigan
Computer Science and Engineering
2260 Hayward Street
Ann Arbor, MI 48109, USA

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

1. Introduction

A clear working definition of anything must relate to a well-defined referent. Definitions of *artificial intelligence* tend to be confusing when they fail to distinguish between two common referents of this phrase. The first common usage, which we will call AI1, refers to the quality of intelligence in some man-made system. The second common usage, AI2, refers to the field of study which addresses systems of the AI1 sort. Thus a definition of AI2 depends on defining AI1, and a definition of AI1 depends on how we define *intelligence*. This commentary will focus on defining intelligence, and then how it relates to AI1 and AI2.

2. Wang's definition of intelligence

Pei Wang, in his paper entitled "On Defining Artificial Intelligence" (Wang, 2019), gives the following proposed definition:

> "Intelligence is the capacity of an information-processing system to adapt to its environment while operating with insufficient knowledge and resources."

This, of course, is a definition of *intelligence* in general, that could apply to humans, animals, or man-made systems, not a definition of AI *per se*. Since the paper talks about a wide range of things that could be intelligent, it is curious that this definition is centered on "an information-processing system." Technically, this term can be considered general enough to cover the whole range of systems Wang discusses, but its common usage tends to imply a computer system. This contradicts Wang's first requirement for a definition, that it have "similarity to the explicandum." It would be better to use a term without computer science implications, such as "agent."

Wang's definition than talks about a system's "capacity ...to adapt to its environment." Certainly a capacity to adapt is an important part of intelligence, but it is not the only important part. Before adapting, it seems an agent would need to act on a moment-to-moment basis in its environment. Perhaps "while operating" suggests ongoing action, but it seems a weak way to say it. Adaptation should not be just to the environment, but also to the agent's own internal needs and goals, which can change over time, as in the development of a child.

Wang says the system must operate "with insufficient knowledge and resources." Later in the paper he expands on this concept, calling it "the *Assumption of Insufficient Knowledge and Resources* (AIKR)." What does it mean that they are "insufficient?" Certainly any finite agent that must operate in real time will have limits on its knowledge and resources. It would be better to talk about "limits" on knowledge and resources, since "insufficient" can only be defined relative to some task, some environment, and some performance measure. What is insufficient for one task may be perfectly sufficient for another, and intelligence should be defined over a wide range of tasks.

3. An alternative definition of intelligence

Consider, then, an alternative approach based on the idea of an *agent*. Key elements of an agent are that it is situated in an environment, that it has limited knowledge, memory, computational capacity, and abilities to perceive and act in its environment, that it chooses its actions moment-to-moment, and that it has goals. An agent may have multiple goals simultaneously, and the goals, the environment, and the agent's capacities and abilities may evolve over time. We can call an agent *intelligent* if its moment-to-moment choices do, over time, lead it toward its goals, and if over time it can learn and adapt by increasing its knowledge and its ability to choose actions that lead it toward its goals.

Given this concept of an agent and what it means for an agent to be intelligent, we offer the following alternative definition of *intelligence*:

> "Intelligence is the ability of an agent, whether human, animal, artificial, or something else, to act in its environment in real time, using its limited knowledge, memory, computational power, and perception and action capabilities, choosing actions at each moment that move it toward its current goals, and to adapt over time by improving this ability to act."

Central to this idea of intelligence is that an agent makes choices on a moment-to-moment basis, that the abilities and capacities to make these choices are limited, and that choices are made to move the agent in the direction of its goals. We would specifically exclude from being intelligent agents computer programs that always produce a certain predetermined output for a given input, systems whose only actions are to categorize the current input, even if the categorization was learned, or systems whose output is primarily determined by some random process that is independent of any perception of the environment.

4. Defining artificial intelligence

Given this definition of intelligence, we can move on to define *artificial intelligence*, in its two senses. In the AI1 sense, it is easy to say that artificial intelligence is the quality of intelligence in a man-made system. In its AI2 sense, AI is the field of study which considers how to design, construct, and evaluate AI1 systems. Now consider how to relate these definitions to some of the ideas in the field.

Consider the question of what constitutes an intelligent artificial agent. Silver et al. (2017) claim that the program they call AlphaGo Zero achieves "superhuman performance" starting "*tabula rasa*" with no human input. However, Marcus (2018) points out some problems with this claim. The deep neural network may learn without labeled input data, but this is only a small part of the whole system. Other parts of the complete agent were hand crafted by human experts, so the agent's performance as a whole actually depends on encoding a lot of human expertise. Thus if we consider an entire agent that acts intelligently in the world, we have a different perspective on AI than when we focus on only a single component, however amazing its performance may be.

5. The Field of AI

From this perspective the question arises of what is the appropriate relationship between the field of AI, AI2, and the study of human intelligence. A prominent textbook (Russell and Norvig, 2010) begins on page 2 with a diagram showing that AI could involve thinking and acting "humanly" as well as thinking and acting "rationally." After further defining these terms, on page 5 they say that the rest of the book will focus just on the "acting rationally" quadrant, dismissing any consideration of modeling human intelligence. Their definition of rationality says that it is "an *ideal* performance measure," thus dismissing from the outset any consideration of human intelligence or the kinds of limitations we have included in our definitions.

Although not all researchers in the field will agree with this approach, it does exemplify the fact that much of AI research today ignores both human intelligence and Herb Simon's (Simon, 1996) concept of "bounded rationality," which takes into account limitations on knowledge and resources. Wang's emphasis on AIKR makes a very important point.

Laird, Lebiere, and Rosenbloom (2017) suggest a different approach. They discuss the concept of "*humanlike minds*," and even propose a "standard model of the mind" based on many years of research in cognitive architectures, in turn informed by research in psychology and cognitive science. Their approach to AI1 exemplifies a part of AI2 that does explicitly consider agency, human cognition, and limits or bounds on rationality. Such an approach fits much better with the definition of artificial intelligence we propose here. Since humans are the only instantiation of full general artificial intelligence we know of today, it seems wise to consider an understanding of human intelligence as we search for better ways of creating artificial intelligence.

References

Laird, J. E.; Lebiere, C.; and Rosenbloom, P. S. 2017. A Standard Model of the Mind: Toward a Common Computational Framework across Artificial Intelligence, Cognitive Science, Neuroscience, and Robotics. *AI Magazine* 38(4):13.

Marcus, G. 2018. Deep Learning: A Critical Appraisal. http://arxiv.org/abs/1801.00631.

Russell, S., and Norvig, P. 2010. *Artificial Intelligence: A Modern Approach.* Pearson Education Limited, third edition.

Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; Chen, Y.; Lillicrap, T.; Hui, F.; Sifre, L.; van den Driessche, G.; Graepel, T.; and Hassabis, D. 2017. Mastering the game of Go without human knowledge. *Nature* 550(7676):354–359.

Simon, H. A. 1996. *The Sciences of the Artificial.* Cambridge, MA: The MIT Press, third edition.

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