sciendo This is an open access article licensed under the Creative Commons BY-NC-ND License.

Journal of Artificial General Intelligence 11(2) 66-67, 2020
DOI: 10.2478/jagi-2020-0003

Submitted 2019-09-27
Accepted 2019-11-04

John McCarthy's Definition of Intelligence

Richard S. Sutton RSUTTON@UALBERTA.CA
University of Alberta
Edmonton, Alberta, Canada

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

Pei Wang (2019), in the target paper, is right to stress the importance of a scientific field having a generally agreed on definition of its subject matter. He is also right when he says that many artificial intelligence (AI) researchers accept, in their textbooks and public statements, that there is no satisfactory way to define intelligence. However, for other AI researchers—including me—this is not acceptable. A field needs to be able to reason, at least in a general way, from a clear statement of its subject matter.

But is there really no standard definition of intelligence within AI? Actually, it is not hard to find a public statement by a prominent AI researcher defining intelligence. The definition given by John McCarthy (1997), the AI researcher who coined the phrase "artificial intelligence," is:

> "Intelligence is the computational part of the ability to achieve goals in the world."

I find this simple and commonsense definition to be useful and satisfying, although it is not specifically mentioned in the target paper.

According to McCarthy's definition, intelligence is an ability, and so of course a system may possess that ability to various degrees. Thus the definition does not make an absolute distinction between systems that are intelligent and those that are not. A person, a thermostat, a chess-playing program, and a corporation all achieve goals to various degrees and thus can be thought of as intelligent to those degrees. This is just as it should be, in my opinion.

McCarthy's definition also specifies that intelligence is the computational part of that ability, ruling out, for example, systems that achieve their goals merely by being physically strong, or by having superior sense organs.

At the heart of McCarthy's definition is the notion of "achieving goals." This notion is clear, but informal. What does it mean, exactly, to have a goal? How can I tell if a system really has a goal rather than just appears to? These questions seem deep and confusing until you realize that a system having a goal or not, despite the language, is not really a property of the system at all. It is a property of *the relationship between the system and an observer*. It is a 'stance' that the observer takes with respect to the system (Dennett, 1989). The relationship between the system and an observer that makes it a *goal-seeking* system is that the system is most usefully understood (i.e., predicted or controlled) by the observer in terms of the system's *outcomes* rather than in terms of its *mechanisms*.

For example, for a home owner, a thermostat is most usefully understood in terms of its keeping the temperature constant—an outcome—and thus for the home owner *the thermostat has a goal*. But for a repairman fixing a thermostat, it is more useful to understand the thermostat at a more

mechanistic level—and thus for the repairman *the thermostat does not have a goal*. The thermostat either does or does not have a goal depending on the observer, depending on whether the outcome view or the mechanism view of the thermostat is more useful. Even for a single observer, which view is more useful may change over time, and thus the same system may change from not having a goal to having one (or vice versa), as when the thermostat repairman fixes his own home's thermostat using the mechanism view, and then uses the thermostat to control the temperature of his house using the outcome view. And of course there may be degrees to which the two views are useful, and thus degrees of goal-seeking-ness. As in the case of intelligence itself, the notion of having a goal or not is not an absolute dichotomy, but a question of degree.

Another good example of goal-seeking-ness varying with the observer is that of a computer chess program. Suppose I am playing the program repeatedly. If I don't know how it works and it plays better than I, then my best understanding of the program is probably that it has the goal of beating me, of checkmating my king. That would be a good way of predicting the near-inevitable outcome of the games, despite how I might struggle. But if I wrote the chess program (and it does not look too deep), then I have an alternative mechanistic way of understanding it that may be more useful for predicting it (and for beating it).

Putting the two ideas together, we can define intelligence concisely and precisely:

> "Intelligence is the computational part of the ability to achieve goals. A goal achieving system is one that is more usefully understood in terms of outcomes than in terms of mechanisms."

References

Dennett, D. C. 1989. *The Intentional Stance*. MIT press.

McCarthy, J. 1997. What is Artificial Intelligence? Available electronically at `http://www-formal.stanford.edu/jmc/whatisai/whatisai.html`.

Wang, P. 2019. On Defining Artificial Intelligence. *Journal of Artificial General Intelligence* 10(2):1–37.