What is an agent?

Anna Harutyunyan

June 2020

There was a paper published in *eLife* last year titled *"Philosophical bias is the one bias that science cannot avoid"*. The authors argue that all science requires philosophical assumptions, but some of them are deliberately examined and chosen, while others are completely implicit. This latter category constitutes philosophical bias.

An agent is an intuitively appealing entity central to artificial intelligence. But it really is only a metaphor – there are no agents in the universe we can observe and measure. This article will meander through the philosophical biases that contribute to our thinking about this metaphor.

Agents

Anything that perceives an environment with sensors and acts upon the environment with effectors is an agent. Human agents perceive with their eyes, ears, and other organs, and act with their hands, legs, mouth, and other body parts. A robotic agent perceives via cameras, and acts with its motors. A software agent perceives and acts with encoded bit strings. The word "agent" itself is suited for this definition perfectly. Originating from the Latin root *ag*, meaning "to act", it literally means "the one who acts". Note then that purely perceptive systems should not be considered agents.

The inner machinery used by the agent to decide how to act is nowadays referred to as a *representation*. A particular representation (like a deep neural network, or its structure) determines the agent's capabilities. The term is related to the notion of *mental* representations, central to cognitive science, and going all the way back to antiquity. A mental representation is presumed to produce mental objects corresponding to physical objects and manipulate them, thus encompassing all perception, memory, and thought. The dualism inherent to this *representationalist* view is associated with René Descartes (1641), and underlies our views on agents.

The intelligent agent concept on its own does not entail a notion of improvement over time, or *learning*. That this is possible at all in a real sense, was a controversial idea in the early days of computing. For example, Ada Lovelace in her detailed account of Babbage's Analytical Engine maintained that:

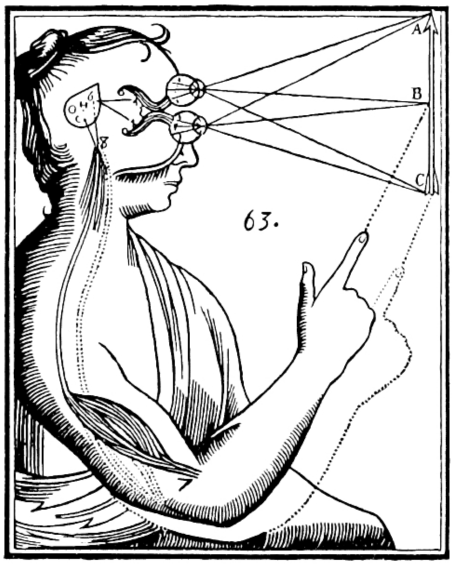

Figure 1: Descartes’s illustration of dualism. Inputs are passed on by the sensory organs to the brain and from there to the immaterial spirit.

> *‘The Analytical Engine has no pretensions to* originate *anything. It can do* whatever we know how to order it *to perform.’*
>
> (Ada Lovelace)

Alan Turing in his visionary *“Computing Machinery and Intelligence”* (1950) collected and argued against nine such objections, and provided some of the first thoughts on what a *learning machine* would look like. Instead of focusing on a baggaged question of whether machines can “think”, Turing asks to consider the *behavior* that would occur if they did, and in particular whether it can be indistinguishable from humans – entities that uncontroversially do think. Without directly rejecting dualism, Turing thus turned to the perspective of *behaviorism*, championed by the likes of B.F. Skinner and Edward Thorndike, which is concerned with analyzing observable behavior rather than mental states and thought.

Modern agents exist on a spectrum that, on the one hand, attempts to explicitly model structures presumed to be contained in mental representations (beliefs, knowledge, memory, intentions, etc), and on the other, following the behaviorist tradition, is satisfied with achieving desired behavior.

History of intelligent agents so far contains a key shift from considering the procedure of acting (or policies) to the procedure of learning (or algorithms) as the central problem. Indeed, we differentiate between agents based on how they learn and not based on how they act.[1] While it is debatable whether we are

[1] Impala and R2D2 are different agents, even if they induce the same behavior.

after policies or algorithms in general, it is the fact that policies are *induced by* learning algorithms that allows us to definitively refute Lady Lovelace's objection.

The agent-environment boundary

> *'The 'skin of an onion' analogy is also helpful. In considering the functions of the mind or the brain we find certain operations which we can explain in purely mechanical terms. This we say does not correspond to the real mind: it is a sort of skin which we must strip off if we are to find the real mind. But then in what remains we find a further skin to be stripped off, and so on. Proceeding in this way do we ever come to the 'real' mind, or do we eventually come to the skin which has nothing in it?'* (Alan Turing, *Computing Machinery and Intelligence*)

We have now outlined the familiar agent-environment model, with the representation serving as a filter between the two entities. In the rest of the article let us challenge this view from several angles.

Outline boundaries

> *'What is the outline? ... It is not something definite. It is not, believe it or not, that every object has a line around it! There is no such line.'* (Feynman et al., *Feynman Lectures on Physics*)

It may seem that the action - perception line, where the environment begins and the agent ends, is definitively drawn, but upon examination this clarity falls away quite quickly. Do humans act with brain synapses with their bodies as part of the environment, or as physical bodies acting upon the external environment as we described earlier? Do contact lenses or prosthetic devices extend the agent or belong to the environment? Which entity pre-processes frames of Atari or makes the agent repeat its actions four times?

The answer of course is – it depends. Each of these boundaries define valid agents. Their choice is not, however, inconsequential for concrete technical purposes. Indeed, Nan Jiang in his recent paper *"On Value Functions and the Agent-Environment Boundary"* (2019) showed that while our algorithms are boundary-invariant, their analyses are not.

The ambiguity of the agent-environment boundary directly relates to a similar ambiguity in quantum physics: Where does the observation apparatus end and the observer begin? Does the apparatus include the computer that interprets its measurements? What about the printer? What about the scientist reading them? Niels Bohr in his radical philosophy-physics considers this, and calls into question the clarity of subject-object dualism provided by Cartesian epistemology. In an early essay (1930), he mentions an example of a man with a stick in a dark room. If the man holds his stick loosely, it is an object, a part of the environment. But if he holds it tightly, it becomes part of the subject, a means of perception. Bohr then insisted that the boundary between the apparatus and its object is *enacted* rather than inherent.

A particular boundary produces an agent that is well-defined, but what is the relationship between agency as a larger concept and the enactment of the boundary? Karen Barad in her ambitious book *"Meeting the Universe Halfway"* (2007) proposes a model of *intra*-action, as opposed to *inter*action between observer and observation. In it, an agent is a boundary-making instrument (as opposed to something induced by a boundary), and the problem focus shifts from entities to phenomena.

Imagining practical intra-action agent-environment models is not directly intuitive. But it is worth remembering that our existing models are arguably only "correct" insofar as they are validated by behaviorism. Interpreting agency through agent-environment boundaries then may be an intriguing next step on the path from policies to learning algorithms.

Mind / matter boundaries

The outline boundary ambiguity applies to all agents. For agents that we consider to be models of general intelligence, however, we can also come back to the other dualism at the core to representaionalism. While Bohr rejected the definitiveness of the apparatus boundary, he, like Descartes, still subscribed to the *humanist* view in which there is a fundamental distinction between the knower and the known, the material and the discursive, mind and matter. The agent is made of different "stuff" from the environment, its representation serving as a mediating filter that translates physical objects into mental objects.

Performatism, also first proposed by Barad (2007), rejects this distinction, and maintains that thought, observation, and inference are as real of practices of engaging with the world as physical action. If observation is action, purely perceptive machines can be considered agents with no contradiction.

Since representationalism is only a perspective and not a fact, it is completely viable to entertain a performative approach. This evokes interesting questions of unifying the interfaces of inference and action.

Embedded agency

> *'Reality is bigger than us.'*
>
> (Ian Hacking, *Representing and Intervening*)

There is another, related angle of viewing the boundary question that has been discussed in artificial intelligence safety and ethics research. The conventional reinforcement learning model views the agent as something placed outside of the environment, optimizing the external signals it receives. The agent's representation is not a part of the environment and is entirely separate. This e.g.

allows for Bayesian agents that, in the extreme, are able to imagine all possible configurations of the world and choose between them rationally.

Embedded agency suggests a different view. The agent and its representation are a part of the environment. This (appealingly) reframes the problem of optimization into a problem of self-improvement, and allows for agents that can reason about themselves. It also reverses the Bayesian perspective: the agent is smaller than the environment (because it is a part of it), and cannot represent the environment perfectly. This is the view we typically take in reinforcement learning. An environment that contains the agent may also contain other similar agents that the agent may learn from. Orseau and Ring in their paper *"Space-Time Embedded Intelligence"* (2012) provide the first formal model of embedded agency.

A pessimist might think that we have merely confused and complicated our understanding of agents, but a realist would see that we have relaxed it. An agent (like an apparatus) is merely a model, a construct, a filter through which some cross-section of reality is reflected. For any particular problem, there is no underlying "correct" agent, but there are many agents that are equally valid. Whether they are equally *useful* is another question, one that is grounded in behaviorism.

When considering agency as a model of general intelligence, however, it may be worth reflecting what the precise meaning of it is, and how our own biases contribute to that reflection. Without an objective underpinning of where the boundary should be drawn, it becomes a choice that itself can be viewed as agency. Turing's 'skin of an onion' analogy then may be read to imply that the structure of intelligence resides not in some presumed core, the "real mind", but in the layering itself.

Acknowledgments. Big thanks to Tom Schaul and Hado van Hasselt for feedback on earlier drafts, and their years of thought put into agents, and Sasha Vezhnevets for pointing me to the perfect book at the perfect time.