On Automata, Autonomy, and How They Relate to Artificial Intelligence

Tim Wang

HPS255 Final Paper, Topic 2, Dec 14th 2020

In our modern intuition, autonomy is often thought to be the basis of human agency, human rights, and even moral responsibility. Autonomy is thought of as the freedom and independence of actions, thoughts and desires. Although it seems clear what autonomy means for humans, it is not obvious what autonomy means for artificial intelligence (AI). Is the notion of autonomy the same for both humans and AIs? If it is, then why do our intuitions suggest otherwise? This is likely because our notion of autonomy has many layers of complexity. As a result, the notion of autonomy may be same for both humans and AI, but the use of autonomy differs for humans and AIs.

The notion of autonomy is closely connected to the concept of automata. To understand autonomy, we should first draw the distinctions and connections between automata and autonomy within their historical contexts. The term 'automaton' comes from the Greek word 'automatos', with roots 'autos' (self) and 'matos' (thinking, willing, or animated). However, this etymology contrasts with our intuitions that automata are "self-moving" things that do not think or have will. This inconsistency likely arose from the influence of the Cartesian view.

Rene Descartes (1596-1650) proposed that automata are mechanisms who are "self-moving". Descartes is known for his dualistic ontology – the view that the world is populated by two substances: matter and mind. According to this view, purely material things like rocks, clocks, and animals are subjected to material and deterministic causal properties – given a cause A (e.g. a clock runs out of battery), a definitive effect B will follow (e.g. the clock stops ticking). In contrast, humans are not limited to the same deterministic causal properties, and this is because humans also possess the substance of mind – a source of free will and unpredictable possibilities. Thus, sometimes we can choose or control our actions, which makes humans indeterministic (or at least not purely deterministic). However, it seemed odd for Descartes to categorize self-moving things like clocks and animals in the same category as non self-moving material things, such as rocks and grass. Thus, the concept of Cartesian automata was created, and Cartesian automata refers to the subset of mechanisms that moves by itself, but is subjected to deterministic causal properties.

Although dualism is unaccepted nowadays, our intuitive understanding of automata was undoubtably shaped by the notion of Cartesian automata. In the modern view, automata are simply things that move by itself, or self-moving things for short.

Autonomy, on the other hand, is linked to thinking and willing. In particular, the term 'autonomy' originates from the Greek word 'autonomos', with roots 'autos' (self) and 'nomos' (law) (Antsaklis and Arash, 2018, p. 24). This can be directly translated to 'self-governing'. In the cartesian view, automata are considered fundamentally different from autonomous beings, as automata do not have minds. Although the cartesian view still serves as basis on discussion related to automata and autonomy, it has been largely replaced by the notion of autonomy, outlined as having the ability and authority for self-government.

The main difference between the cartesian view and the modern view is that the modern view allows the possibility for self-moving automata to also be autonomous. For instance, numerous AI technologies, such as self driving vehicles, are considered to be autonomous when navigating their immediate surroundings, as they can avoid obstacles and get themselves unstuck from tight corners without any external intervention. On the contrary, not all automata should be considered autonomous. For instance, a simple clock is usually not considered as an autonomous machine. Although a clock is self-moving, and can move its arms without external intervention, it does not have the choice or the power to do anything besides displaying the time that it is mechanically "instructed" to display. The self driving cars are different from clocks in the sense that they have a choice to an extent. It has at least two choices when avoiding an obstacle – to move to the left or to the right. This shows that perhaps the cartesian view is still embedded into the modern view of autonomy, as things that have a deterministic nature are not considered autonomous in the cartesian view. Subsequently, we may place emphasis on the element of control when it comes to autonomous entities, such that to be autonomous impinges on having a choice in certain situations.

With the current view of autonomy explained, one may see that the notion of autonomy is the same for humans and AIs. The modern view is consistent with our intuitions of human autonomy of freedom and independence of actions, thoughts and desires, and is also consistent with AI developers' or programmers' notions of autonomous machines, as evidenced by the self driving vehicle example. However, does this suggest that autonomy for humans means the same

for AIs? Once again this is at odds with our intuitions. Namely, we do not think that current AIs can have freedom and independence of actions, thoughts and desires. This is not because we have unethical AI laws that forbids them from having freedom, but because current AIs cannot think and desire in the same sense as humans do. In brief, it is easy to see how autonomy for humans and AIs both fall under the broad scope of the same notion in the modern view while there are still some differences between the two. These differences lie within the use of autonomy.

To find the difference between the two uses of autonomy, one must explore the concept of autonomy in further depth. One thing to note is that autonomy can have multiple dimensions. As suggested earlier, self driving vehicles are considered to have autonomy when navigating their immediate surroundings. One exemplar of this is a self driving delivery robot going by the name of Kiwibot, which are already in service in at least 10 collages in the United States (Coldewey, 2019). These delivery robots are about knee-high, and have locking and insulating compartments that can hold a customer's order within. During the delivery process, the robot navigates most sidewalks and avoids pedestrians without external help. Although comfortable navigating its immediate surroundings, these robots do not know how to reach the customers who placed delivery orders. Thus, human workers are added "in the loop", monitoring the robot and updating its waypoint every five seconds on average. Developers call this an example of "semi-autonomous" robots, as they clearly need external help navigating to the destination, despite the fact that they are fully autonomous when navigating its surroundings. From this, one might distinguish that there are at least two dimensions of autonomy: a material dimension, and an informational dimension. The material dimension of autonomy is predicated on the ability for an agent to move or operate without external intervention. In contrast, the informational dimension of autonomy is predicated on knowledge in a sense of knowing how and knowing that. Thus, one may say that the Kiwibot possesses autonomy in the material dimension, but lacks autonomy in the informational dimension.

Knowledge, or the informational dimension of autonomy, plays a major role in autonomous behaviors. Autonomy involves making decisions without external intervention, but is it possible to do so without knowledge about those decisions? Imagine making a decision on whether to walk the dog before knowing what 'walk' and 'dog' even mean. It is impossible to make autonomous decisions in this case without knowing the concepts of 'dog' and 'walking'. There are countless examples from current AI systems that either fail or have difficulties performing information

processing tasks. A notable example is evidenced by the website now known as the Amazon Mechanical Turk (MTurk), which once belonged to a project attempting to use AI systems to process general information and verify data. This project ended in failure like many other information processing AIs did. Instead of computers, now human workers who respond to requesters perform these tasks of labelling images for machine learning training, finding duplicate items in data sets, verifying information and moderating content. One example of such task is training a computer to recognize pictures of cats with the strings c-a-t. These tasks seem trivial for us humans, yet not for AI systems. This is the main reason why the use of autonomy for AIs differ from that of humans. AI systems struggle with connecting references and knowing basically what is what, while we seem to connect references and use them to make autonomous decisions effortlessly. In short, one should recognize that knowing how and knowing that are essential to autonomous behaviors in the informational dimension. The information gap between current AI systems and humans is the main factor that results in the different use of autonomy for humans and AIs.

Autonomy not only has multiple dimensions, but different levels and different degrees of autonomy within each dimension. This is easy to demonstrate. Within the material dimension of autonomy, those who have larger variety of available movements can be thought of as having a higher level of material autonomy. For instance, a space robot that can navigate rough surfaces and climb steep mountains have more material autonomy than the Kiwibots that can only navigate smoothly paved sidewalks. While within the informational dimension, those who have a larger knowledge base than others can be interpreted as having a higher level of informational autonomy. For instance, the impressive chess playing AI, AlphaGo, can calculate results many steps ahead of the current one. This suggest that AlphaGo possess more informational autonomy than most other AIs and most humans in the domain of chess playing. However, these levels of autonomy must be compared within their respective domains. It would be odd to claim that cheetahs (the fastest land animal in our knowledge) have more material autonomy than gold fish. This is because the domain of traveling is different for cheetahs (who travel on land) and gold fish (who travel in water), and thus incompatible for this kind of comparison. Thus, this type of comparison between different domains should be considered as different degrees of autonomy, and they cannot be compared with levels. In the example of cheetahs and gold fish, one can only conclude they have a different degree of autonomy within the same material dimension.

Having established that autonomy has at least three layers of complexity – in different dimensions, levels within a dimension, and degrees within a dimension – there is yet another layer of complexity to be considered: the integration between the dimensions. Although having the distinction between different dimensions of autonomy is useful for seeing the different uses of autonomy, the material and informational dimension often overlaps. For human autonomy, it is obvious: Although some have more autonomy than others, most of us have both material autonomy while navigating through physical environments and informational autonomy while making decisions. The intersection between dimensions of autonomy is not always easy to see. For instance, the Kiwibot in the example before was categorized as having material autonomy while it lacking informational autonomy. However, sometimes it is not possible to only possess one dimension of autonomy. In order to travel around, the Kiwibot must possess at least a limited sense of informational autonomy. It needs to be able to move forward, make left and right turns, and back up, but it also need to know how to combine these movement in order to avoid obstacles or get itself unstuck from tight corners. This shows that the Kiwibot does possess a limited sense of informational autonomy, although most of the autonomy ascribed to it is the material dimension.

To summarize, there are many layers of complexity to the notion of autonomy. Namely, there are at least two dimensions – the material dimension and the informational dimension – and these dimensions often overlap. Within each dimension, there are different levels and degrees of autonomy. These complexities are the reason why the notion of autonomy appears different for humans than it does for AIs. However, it is not the notion of autonomy, but the use of autonomy that is actually different for humans and AIs. This is because most current AI technologies struggle to have high levels of informational autonomy. In contrast, humans possess high levels of both material and informational autonomy. Thus, the use of autonomy for current AI technologies refers to a lesser extent of human autonomy. This is not to say AIs will always have lower levels or a lesser extent of autonomy than humans. When AIs can rival the intelligence of their human creators, they may possess levels of autonomy even higher than what we can imagine today.

References

Antsaklis, Panos J., and Arash Rahnama. "Control and Machine Intelligence for System Autonomy." *Journal of intelligent & robotic systems* 91.1 (2018): 23-34. Web.

Coldewey, Devin. “Kiwi's Food Delivery Bots Are Rolling out to 12 More Colleges.” *TechCrunch*, TechCrunch, 25 Apr. 2019, techcrunch.com/2019/04/25/kiwis-food-delivery-bots-are-rolling-out-to-12-new-colleges/?guccounter=1.

Hall, Jessie. “Week 7 lecture video 2: Automata, Autonomy, and Mechanisms” Course content for *HPS255: History and Philosophy of Artificial Intelligence*, University of Toronto.