The age of machine autonomy?

Author: Thomas Christian Bächle

Can machines be autonomous – or is it a human prerogative? This categorical question dominates many discussions on our relationship to purportedly intelligent machines. A human vs. machine rhetoric, however, does not get us very far. On the contrary, it even sidetracks us and distracts our attention from the more important issue of how these autonomous systems change the ways we humans relate to the world at large and to each other. In a much broader sense, robots – as simulated humans in particular – are bound to shake up the fundamentals of what we regard as human autonomy.

The largely useless either-or understanding of autonomy

"Autonomous" has become a commonplace characterisation of computer systems that are capable of operating in an unforeseeable manner. In popular discourse, but also many academic ones, this attribute is often a stand-in for anything related to artificial intelligence. As a consequence, the usual vagueness of the AI label is in many instances traded in for that of another.

Despite this blurriness, the term "autonomous" comes with some advantages in many of the current debates on machine autonomy. First of all, it hints at a distinct feature of AI-related computer systems, which sets them apart from technologies of the past: an astounding – and ever-growing – independence from human control, intervention and agency. In engineering or computer science the autonomous features of a system are understood as the next step from automation, which was largely about preconceiving all possibilities and eventualities in a system's programming. Automatic systems react to a simple threshold value; by contrast, automated systems follow, in an elaborate but still pretty narrow sense, the rules that are predetermined in detailed and intricate human-made recipes, in effect a top-down model in action. Autonomous systems, on the other hand, are able to abstract patterns from unknown data (a bottom-up approach; this is their machine learning component) and are equipped in their programming to adapt to that which has not been preconceived step-by-step, enacting much higher degrees of agency compared to their automatic companions (see e.g. Scharre 2018, 26–34).

Second, and besides these technologically-grounded meanings, the label "autonomous" also has a very practical advantage: People instantly know – or think they instantly know – what is meant by it. The label is popular, evokes well-known images of self-driving cars or robots and is at the same time highly inclusive. In this sense, it is capable of facilitating much needed exchange between the broader public, political decision-makers and expert discourses and also between academic disciplines; the term spans technical, social and philosophical approaches. This partly explains its popularity, well-established somewhere besides the even more generic

"AI" speak.

It is hardly a surprise that the characteristics of this general and inclusive term "autonomy" lead to problems in other areas. While the technical disciplines, such as engineering and computer science, have managed to narrow down the meaning of autonomous systems considerably and establish a pretty circumscribed interpretation (by and large the one sketched above), the terminology is constantly at odds with approaches that are rooted in philosophical interpretations of autonomy, not least because their main point of reference is not a technological entity but the human being. The very idea of human autonomy implies that a distinct philosophical or anthropological discipline is (or at least should be) in charge, one that seeks to explore the fundamentals of the human condition, with a long-standing conceptual tradition deeply rooted in European thought. It evokes connotations with figurations such as free will and the rational subject, which more often than not deliberately and firmly position the meaning of autonomy against its application in technological discourses.

While being inclusive, this term-for-all comes with some serious contradictions: it may be the same concept, but we usually mean very different things, sometimes even arguing from irreconcilably different positions. These positions become mutually exclusive, when 'autonomous' is articulated as a marker of what is human. And sometimes (in a quite foreseeable and self-evident manner), the position comes down to merely arguing that machines are not really intelligent, do not have free will and hence cannot be autonomous at all.

The disciplinary territories reveal their respective limitations. This is also the reason why an argument that seeks to explore the differences between human and machine autonomy may be perfectly viable as an exercise in debating – but will still lead us nowhere. When the respective premises are so different from the outset, some of the answers are quite predictable: Machines are not autonomous in the human sense, but it is also questionable whether humans are either.

Ironically, the fundamental issue with a positive understanding of autonomy (presupposing free will of the rational subject) points at a predicament that both machines and humans have very much in common regarding their autonomous qualities and capabilities: Both of 'their autonomies' not only heavily rely on external factors; these external factors constitute the very idea of autonomy in practice, which relies on a framework of cultural techniques, symbolic structures, technology, down to somatic conditions and needs.

So let's look at this relationship more closely.

Modelling autonomy: a relation, not an essence

Most debates on the topic of autonomy and autonomous systems can (in an admittedly simplified manner) be divided along the lines of two major interpretations of autonomy. The first impulse is to tie it to figurations of the human, such as the rational and free subject, reasoning or a human's innate qualities and values (see Christman 2020 for an overview). The rational

subject makes informed decisions and exercises free will. In the most obvious sense, it can easily be argued that this traditional understanding is challenged by AI when autonomous systems partly take over elements of decision-making processes that were once the prerogative of humans. It can also be argued that the opacity of said systems challenges the very foundation of human autonomy, as we can no longer be completely sure what human decisions are based on. Or, in very practical terms: How can we make informed decisions that express our intentions when the options we choose from are preselected and hence predetermined by a technical 'AI' system? Also in line with this first interpretation of autonomy, but a little more subtly, the independence of autonomous systems forces us humans to question the very idea of our own human agency. In general, this essentialist understanding of autonomy, in public or science discourses alike, often serves to rhetorically and conceptually portray machine agency as something categorically different, an essence of the human condition, so to speak.

The second understanding of autonomy, on the other hand, regards it as a relational quality that can only exist within specific symbolic or material frameworks. In this line of thought, it is these frameworks of autonomy that create complex entanglements of human/machine autonomy in the first place. In other words, there is no autonomous subject or agent per se; autonomy is rather to be seen as an effect of relations that enable autonomous processes. In more specific terms, this translates into structures that should be seen as the conditions of autonomy, which also reflects its paradoxical character: Individual freedoms need to be curtailed, for example by rules set in laws or norms, in order to safeguard individual autonomy, presenting as a dynamic dialectic of normativity and freedom (Khurana 2013). We rely on these social structures, as well as language or our bodies, in order to 'be' autonomous (cf. Rössler 2021).

These relational principles are also reflected in both the conceptualisation of autonomous systems themselves and the human/machine relationship that informs this conceptualisation. In a basic sense, the concept 'autonomous' – in both humans and computer systems alike – denotes two meanings, namely "self-sufficiency — the capability of an entity to take care of itself" and "self-directedness, or freedom from outside control" (Bradshaw et al. 2012, 3). As we always deal with relations, though, "strictly speaking", the term "autonomous systems" is a "misnomer" (Bradshaw et al. 2012, 5), a logical impossibility, as no entity is either fully self-sufficient nor self-directed. Autonomy, therefore, as noted above, denotes the system of relations, not the 'computer system' or the human as distinct entities. Autonomy is a quality that is enacted in these human/machine systems, not a property.

This relational aspect of autonomous human/machine systems is also reflected (albeit often implicitly) in the degrees of control over the technical system that are distinguished in an effort to classify them. In this sense, the notation "in, on and out of the loop" is often misunderstood as characterising machine autonomy ex negativo by classifying the possibilities of human intervention. "In-the-loop" usually refers to directly executed control, when an action is initiated, "on-the-loop" refers to systems whose actions can be prevented or aborted and "out-of-the-loop" refers to systems that no longer require human control but are, most of the time, (still)

monitored by human agents. This classification, however, does not so much characterise machine autonomy as it acknowledges the relational understanding of autonomy (Bächle/Bareis 2022).

As this relationship is so central — how can humans intervene?; how do machines shape human behaviour? — the modes of this human/machine interaction play a key role (Suchman 2007). Therefore, AI is not an abstract idea; it is tied to a particular materiality, an interface structure that shapes its place in the world.

Humanoid robots: the more autonomous AI?

Besides these conceptual issues, modelling autonomy is also a very productive practice of meaning-making that not only informs us about necessary debates on the human condition but is also an expression of the (historically constant) socio-cultural desire to develop a technology that mirrors ourselves. This becomes most obvious with robots, which can to varying degrees feature human-like design elements, such as stylised facial expressions on a screen or a voice.

The material dimension of the robot body, however, makes them a unique kind of AI. As we have seen so far, it is the relational aspects that directly determine the idea of autonomy. In both a functional as well as a morphological sense, a humanoid robot features a distinct range of skills (such as the capacity to move around freely) and features (such as a human-like material face) that sets it apart from more abstract notions of AI autonomy that is associated with software, algorithm-based learning systems or an 'intelligent' bot. In other words, a humanoid robot can potentially establish much more elaborate material and social connections to its surroundings.

In a functional sense, a humanoid robot (commonly a bi-pedal system with arms and hands), can show a higher degree of self-directedness that allows it to interact with and move in (unknown) environments. These recursive sensory-actuation loops (the machine equivalent of human perception and action) allow it to automatically refine representations and models of the world 'outside', an approach that is very much in line with the bottom-up understanding of machine learning sketched above. Humans are not required — at least in principle — for the AI system to explore its surroundings.

Closely connected to these features are, in a morphological sense, the human-like elements that serve as an important interface, including gestures, the mimicry of facial expressions or its 'body language'. The humanoid system not only grants access to forms of human agency but also to an important mode of communication that constitutes an important part of social reality, opening the realm of emotions and the affective dimension to human/machine interaction. Humanoid robots have the potential to act as truly social interfaces.

Where does this leave us with our understanding of autonomy and autonomous systems? A much closer interconnectedness between the embodied computer system and its environment significantly widens the range of potential relations: this makes the autonomous humanoid robot with social features the most autonomous artificial agent there is right now. And as for our

traditional understanding of human autonomy, the ideas of perfect simulation and deception become central. We are on the verge of entering a mediated social environment where we can no longer be sure whether we are interacting with a human agent (another sentient being, that is) or an artificial agent, a bot or robot that is devoid of soul and mind. In the face of this increasing doubt, the basis for informed decision-making wavers — and along with it our relations to the world and with it our own autonomy.

References

Bächle TC, Bareis J (2022) "Autonomous weapons" as a geopolitical signifier in a national power play: analysing AI imaginaries in Chinese and US military policies. European Journal of Futures Research 10, 20 (2022). https://doi.org/10.1186/s40309-022-00202-w

Bradshaw J, Hoffman R, Woods D, Johnson M (2013) The seven deadly myths of "Autonomous Systems". IEEE Intelligent Syst 28:54–61.

Christman J (2020) Autonomy in Moral and Political Philosophy, The Stanford Encyclopedia of Philosophy (Fall 2020 Edition), Edward N. Zalta (ed.). https://plato.stanford.edu/archives/fall2020/entries/autonomy-moral/

Khurana T (2013) Paradoxes of autonomy: on the dialectics of Freedom and normativity. Symposium 17(1):50–74. https://doi.org/10.5840/symposium20131714

Rössler B (2021) Autonomy: An essay on the life well-lived, Cambridge.

Scharre P (2018) Army of none: autonomous weapons and the future of war. W. W. Norton & Company, New York.

Suchman L (2007) Human-machine reconfigurations: Plans and situated actions. Cambridge, New York.