Bert Heinrichs

Artificial Intelligence and Rational Discourse

1. Introduction

The term ›artificial intelligence‹ (AI) which John McCarthy invented for the famous Dartmouth Summer Research Project on Artificial Intelligence proposal he submitted to the Rockefeller Foundation in 1955 with his colleagues Marvin Minsky, Nathaniel Rochester, and Claude Shannon, has always been controversial. It might, therefore, seem idle to make it the subject of an examination again. However, the term has been almost omnipresent for several years now and it has significantly shaped debates about many technical developments. It is not only the word ›artificial intelligence‹, but also the divergent meanings, or rather, the promises, that many associate with it that have made it controversial. In this paper, I aim at presenting some thoughts on the notion of artificial intelligence and finally place a second notion—that of rational discourse—alongside the former one. In this way, I hope to introduce a perspective that moves the infamous, but in my view over-emphasized question ›When will machines have surpassed us in X?‹ aside. The more interesting question, I submit, is ›Who is ›us‹?‹.

2. Defining Artificial Intelligence

Some argue that AI is simply the field devoted to building a machine that can pass the notorious Turing test.[1] Though highly controversial, the Turing test certainly is still an important benchmark.[2] For my considerations, however, other attempts to define AI are more relevant in the first place. A frequently read—though deliberately tenta-

[1] Turing 1950.
[2] Moor 2003.

tive—definition of AI is as follows: »Artificial Intelligence is the study of how to make computers do things at which, at the moment, people are better.«[3]

According to this approach, AI research is about human capacities, or at least capacities that humans have, and about implementing them in artificial systems, and possibly even about designing these systems to outperform humans in terms of these capacities. Of course, this is still rather vague and imprecise.

In their influential textbook on AI, Stuart Russell and Peter Norvig distinguish two dimensions in AI definitions that can help get a slightly better and more accurate handle on the matter. Accordingly, definitions can be distinguished along two different types of goals, first, human-based vs. ideal rationality-based definitions, and second, reasoning-based vs. behavior-based definitions.[4] This results in a total of four types of definitions, which apparently capture a large proportion of all proposed definitions of AI. The above-mentioned definition by Rich, Knight and Nair, for example, clearly falls into the category human-based / behavior-based, for it highlights that AI research aims at making computers *do* certain things in which *humans* are currently especially good. In contrast to Rich, Knight and Nair, Russell and Norvig themselves emphasize *perfect rationality*, and therefore fall into the category ideal rationality-based / behavior-based. They state the goal of their approach to AI as designing »successful agents«.[5] Later, they define an ideal rational agent as follows:

For each possible percept sequence, a rational agent should select an action that is expected to maximize its performance measure, given the evidence provided by the percept sequence and whatever built-in knowledge the agent has.[6] From this it should be clear that the authors focus on perfect rationality and behavior.

In their introduction to the *Cambridge Handbook of Artificial Intelligence*, Keith Frankish and William M. Ramsey observe:

> Very generally, artificial Intelligence (AI) is a cross-disciplinary approach to understanding, modeling, and replicating intelligence and

[3] Rich, Knight & Nair 2010, 3.
[4] Russell & Norvig 2010, 4–8.
[5] Russell & Norvig 2010, 34.
[6] Russell & Norvig 2010, 37.

> cognitive processes by invoking various computational, mathematical, logical, mechanical, and even biological principles and devices.[7]

The lack of reference to actions makes this definition fall into the reasoning-based category and the lack of reference to humans puts it into the ideal rationality-based subfield. This shows that there are also definitions belonging to this category in use.

However, the majority of AI research today seems to focus on behavior, i. e. on how to make computers *do* certain things. One may actually assume that it is never exclusively about reasoning processes, so that behavior-based definitions—at least if one takes a broad understanding as a basis—can serve as a comprehensive definition of AI in the one dimension.

This is in line with an expansion of the term ›agent‹ that can be observed in the context of AI research, but also in philosophy of AI. Again, Russell and Norvig are a case in point. In fact, they consider AI as the field concerned with the development of intelligent *agents*.[8] They do so against the background of a very broad understanding of ›agency‹:

> An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators. […] A human agent has eyes, ears, and other organs for sensors and hands, legs, vocal tract, and so on for actuators. A robotic agent might have cameras and infrared range finders for sensors and various motors for actuators. A software agent receives keystrokes, file contents, and network packets as sensory inputs and acts on the environment by displaying on the screen, writing files, and sending network packets.[9]

This understanding of the term ›agent‹ is by no means self-evident. On the contrary, it even contradicts an established way of using it. It therefore seems appropriate to take a closer look at the concepts of agent and agency.

3. Agents

Traditionally, the term ›agency‹ was used as an opposite term to behavior: While non-human animals behave, humans act. In this sense,

---

[7] Frankish & Ramsey 2018, 1.
[8] Russell & Norvig 2010, 4f.
[9] Russell & Norvig 2010, 3.

agency is closely linked to other notions, especially to the notion of responsibility. Thus, the concept of responsibility is often brought into play when it comes to the argument that non-humans (non-human animals and artificial systems alike) cannot be agents, because they cannot assume moral responsibility. However, this conventional understanding of ›agency‹ has been replaced for some time by a much broader and more general understanding.

At a sufficiently high level of abstraction,[10] an agent is someone who (or something that) can act, and to be able to act is to be able to effect changes in the world.[11] As Floridi correctly observes, this abstract definition includes not only ordinary people but also earthquakes since they, apparently, are able to effect changes in the world. Floridi indicates that the definition's level of abstraction is too high for his purposes, and this holds true also for the present purpose. He therefore suggests turning to a lower level of abstraction so that the following criteria are included: interactivity, autonomy, and adaptability. Note that autonomy is not meant as a philosophically rich notion here, but merely designates the ability of a system »to change its state without direct response to interaction: it can perform internal transitions to change its state.«[12]

The crucial consequence of this understanding of the notion of an agent and agency respectively is that it makes it possible to distinguish agency from *moral* agency. While all moral agents are, naturally, agents, not all agents need to be moral agents. As mentioned, this stands in contrast to a long philosophical tradition that conceives of agency exclusively as moral agency. However, it allows entities that share interesting properties to be grouped under one term. If these entities fulfill the three criteria mentioned above, namely interactivity, autonomy, and adaptability, they all belong together and are, in turn, distinct from entities which do not meet the criteria. In this perspective, there are consequently two basic categories: agents and non-agents. If one follows this approach, then it should be uncontroversial that some AI systems are agents—and in fact, this is what Floridi and many scholars from the field of AI argue for. There is nothing wrong with that. In some ways, AI systems and humans share more than AI systems and chairs or humans and chairs respectively. In short, both humans and AI are agents—or at least can be agents.

---

10 Cf. Floridi 2015, Chap. 3.
11 Cf. Floridi 2015, 140.
12 Floridi 2015, 140.

4. Intelligence

If the term ›agent‹ is open to several interpretations, then the term ›intelligence‹ is even more so. In a statement first published in December 1994 in the *Wall Street Journal* and later as an editorial in the journal *Intelligence*, 52 experts suggested the following broad understanding of the concept:

> Intelligence is a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience. It is not merely book learning, a narrow academic skill, or test-taking smarts. Rather, it reflects a broader and deeper capability for comprehending our surroundings—›catching on,‹ ›making sense‹ of things, or ›figuring out‹ what to do.[13]

This definition is primarily directed against racist appropriations of the term and explicitly interprets intelligence as a human capability. However, there is nothing to be said against detaching this general definition from the historical context of its origin and applying it to current developments. If one does so, then one will have to concede that some modern computer systems are quite capable of meeting the requirements described. They are certainly able of comprehending their surroundings, catching on, making sense of things, and figuring out what to do. The same will be said of many animal species. This simply means that there are many entities to which the characterization ›intelligent‹ fits.

What follows from all this? First of all, it makes sense to speak of artificial intelligent agents, i.e., one must concede that there are artificial systems that are agents and that they are intelligent. Conversely, it follows that the question ›Can machines act intelligently?‹ must be answered positively. On this very question Russell and Norvig remark:

> Our definition of AI works well for the engineering problem of finding a good agent, given an architecture. Therefore, we're tempted to end this section right now, answering the title question in the affirmative. But philosophers are interested in the problem of comparing two architectures—human and machine. Furthermore, they have traditionally

[13] Gottfredson 1997, 13.

> posed the question not in terms of maximizing expected utility but rather as, »Can machines think?«[14]

In the end, then, the definitional question of what AI is leads back to a comparison of two architectures, humans, and artificial systems. Even if we concede that both architectures allow intelligent agents to be implemented, there still seems to be an open question.

Russell and Norvig discuss a whole series of philosophical arguments and finally conclude that (phenomenal) consciousness marks a significant difference that continues to be difficult, but not really of central importance from an engineering perspective.

> Running through all the debates about strong AI—the elephant in the debating room, so to speak—is the issue of consciousness. […] Qualia are challenging not just for functionalism but for all of science. Turing himself concedes that the question of consciousness is a difficult one, but denies that it has much relevance to the practice of AI […]. We agree with Turing—we are interested in creating programs that behave intelligently. The additional project of making them conscious is not one that we are equipped to take on, nor one whose success we would be able to determine.[15]

One can summarize this train of thought in the following way: The comparison between artificial systems and humans ultimately boils down to the question Turing rejected as too vague: ›Can machines think?‹ The answer to this question depends on how one deals with the problem of consciousness: If ›thinking‹ is understood to mean intelligent action, then they can; if it is understood to mean conscious action, then they cannot. AI in practice is primarily concerned with the former and not with the latter. But perhaps the question whether machines can think is just a proxy, at least for those who don't take an engineering perspective. The real question might be ›Can machines be like us?‹

If you actually think this question is interesting, then another traditional term might help: rationality. Authors like Russell and Norvig and others use ›intelligent action‹ and ›rationality‹ almost synonymously. They maintain: »A rational agent is one that acts so as to

[14] Russell & Norvig 2010, 1021.
[15] Russell & Norvig 2010, 1033.

achieve the best outcome or, when there is uncertainty, the best expected outcome.«[16]

A broader understanding of rationality as opposed to intelligence is mediated by the notion of reasons or, more precisely, the exchange of reasons. Therefore, I suggest instead of asking ›Can machines act intelligently?‹ (Yes) or ›Can machines think?‹ (Unclear) focusing on the question ›Can machines be rational?‹—if only as an intermediate step to addressing the question, ›Are they like us?‹

5. Rationality as the Giving and Asking for Reasons

Among others, Robert Brandom has proposed and detailed an approach of rationality as the giving and asking for reasons. Following Wilfrid Sellars, Brandom has used the metaphor of the space of reasons for this. Famously, Sellars argued in his *Empiricism and the Philosophy of Mind*:

> The essential point is that in characterizing an episode or a state as that of knowing, we are not giving an empirical description of that episode or state; we are placing it in the logical space of reasons, of justifying and being able to justify what one says.[17]

Brandom subsequently highlighted, among other things, the social structure of the space of reasons. He writes in this regard:

> Thinking of things this way, assessing someone as having successfully achieved the status or standing of a knower involves adopting three different attitudes: *attributing* a commitment, *attributing* an entitlement, and *undertaking* a commitment. There is nothing in principle mysterious about such assessments, nor, therefore, about the standing being assessed. Knowledge is intelligible as a standing in the space of reasons, because and insofar as it is intelligible as a status one can be taken to achieve in the game of giving and asking for reasons. But it is essentially a *social* status because it incorporates and depends on the social difference of perspective between *attributing* a commitment (to another) and *undertaking* a commitment (oneself). If one *individualizes* the space of reasons, forgetting that it is a shared space within which we adopt attitudes towards *each other*—and so does not think about standings in the space of reasons as socially articulated, as potentially

[16] Russell & Norvig 2010, 4.
[17] Sellars 1997, § 36.

> including the social difference of perspective between attributing and undertaking commitments, that is, between your standing and mine—then one will not be able to understand knowledge as a standing in the space of reasons.[18]

If one understands rationality in this way, i. e. in terms of giving and asking for reasons, it is essentially a social practice, or a game played by normative beings, i. e. beings who can undertake commitments.

This brings us back to Russell and Norvig's as well as Floridi's wider notion of agents and agency. It is possible, as I said, to decouple the notion of agency from that of responsibility so that AI systems also fall under it. And it is also possible to understand intelligence in a somewhat neutral sense. What is not possible, according to Brandom, is to decouple the notion of rationality from that of epistemic responsibility, for to be rational is to undertake commitments.

Now, moral responsibility and epistemic responsibility share the feature of being normative. Thus, it is impossible to conceive of rationality in non-normative terms. Moreover, if we widen the notion of agent and agency to the effect that humans and AI systems can fall under it, we end up with a notion that includes both normative and non-normative entities. This is no problem, if putting the two in one category (for certain purposes) does not obscure this important difference. What is most important however is that admitting the existence of moral and non-moral agents is not enough for it can lead to the view that AI is like us in every respect. True, AI systems can be like us in terms of certain capacities, namely those capacities that fall into the category ›intelligence‹, but this is not the same as rationality. AI systems cannot give reasons for what they are doing. Of course, AI is not like us in another respect: we are alive. I leave this aspect out of consideration here, although of course it should be included in a more extensive account of the differences between humans and AI systems.

The upshot of all this is that notions such as ›intelligence‹ and ›agent‹ can be used to denote both humans and AI systems and by using them we pick out interesting features of both humans and AI systems. It would, however, be wrong to conclude that AI systems are almost like us in every respect, except that for the moment they are not ›moral‹. They are not ›normative‹ and this is a much more far-reaching claim, for it implies that they are not ›rational‹.

---

[18] Brandom 1995, 903–904. See also Brandom 1994, 199–206.

6. Who are ›we‹?

In the opening passage of the first chapter of *Making it Explicit* Robert Brandom remarks: »We are the ones for whom reasons are binding, who are subject to the peculiar force of the better reason.«[19] This ›we‹ is not, as Brandom highlights, exclusionary or disparaging and certainly not simply limited to humans. This ›we‹ is basically open for other lifeforms, Martians, and also for artificial intelligent agents. But it says something very fundamental about those who use it—it characterizes them as *normative beings*. Questions concerning AI are, eventually, questions about us as normative beings.

The success story of AI will continue. There will certainly be setbacks and frustrations, too. But the presence of AI in daily life will increase and the performance of systems will improve. One should therefore be very careful with predictions of the kind ›Machines will never be able to do X‹. They have been proven wrong too many times. Above all, it should be clear that the intelligence of artificial systems will grow steadily—at least if intelligence is understood to mean the ability to solve problems in the broadest sense. This should be seen as welcome news and by no means as a horror scenario. AI will make life easier for many people. This is not blind enthusiasm for technology, but a reasonable forecast based on experience with other technologies. Of course, there will also be problems and possibly certain groups of people will be worse off because of the development. It is the task of a far-sighted policy to mitigate these negative consequences.

A completely different question is what influence the further development of AI will have on our self-image. If we see ourselves as beings who can play chess or Go particularly well, then this self-image will suffer considerable damage or has already done so. AI systems are already much better at this and there will be more and more areas where their skills trump ours. Perhaps, however, we should not see ourselves primarily as intelligent beings, but rather as rational beings, and more precisely as beings for whom it is characteristic that they play the discursive game of giving and asking for reasons. We are beings who ask who ›we‹ are and that is not a question that can be answered with intelligence alone. Moreover, it is not a capacity in which one can outperform another.

---

[19] Brandom 1994, 5.

It could be that one day we will find that this ›we‹ also includes other beings than humans—higher animals, inhabitants of other planets, or AI systems. If that should happen, that is, if a representative of one of these groups asks, ›Who are you?‹, then we should probably answer ›Someone like you‹ and welcome them into our midst. Today it is completely unclear whether this will ever happen.

References

Brandom, Robert (1995): Knowledge and the Social Articulation of the Space of Reasons, in *Philosophy and Phenomenological Research* 55, 895–908.

Brandom, Robert (1994): *Making It Explicit. Reasoning, Representing, and Discursive Commitment*, Cambridge, Mass.: Harvard University Press.

Floridi, Luciano (2015): *The Ethics of Information*, Oxford: Oxford University Press.

Frankish, Keith & Ramsey, William M. (2018): Introduction, in Frankish, Keith; Ramsey, William M. (eds.): *The Cambridge Handbook of Artificial Intelligence*, 3rd ed., Cambridge: Cambridge University Press, 1–11.

Gottfredson, Linda S. (1997): Mainstream Science on Intelligence: An Editorial With 52 Signatories, History, and Bibliography, in *Intelligence* 24, 13–23.

Knight, Kevin; Rich, Elaine; Nair, B. (2010): *Artificial Intelligence*, Noida: Tata McGraw Hill.

Moor, James H. (ed.) (2003): *The Turing Test: The Elusive Standard of Artificial Intelligence*, Dordrecht: Kluwer.

Russell, Stuart J. & Norvig, Peter (2010): *Artificial Intelligence. A Modern Approach*, 3rd ed., Upper Saddle River: Prentice Hall.

Sellars, Wilfrid (1997): *Empiricism and the Philosophy of Mind*, Cambridge, Mass.: Harvard University Press.

Turing, Alan M. (1950): Computing Machinery and Intelligence, in *Mind* LIX, 433–460. Reprinted in Copeland, B. Jack (ed.) (2004): *The Essential Turing*, Oxford: Oxford University Press, 441–464.