Philosophy & Technology (2021) 34:1569–1580
https://doi.org/10.1007/s13347-021-00475-2

RESEARCH ARTICLE

Aliens in the Space of Reasons? On the Interaction Between Humans and Artificial Intelligent Agents

Bert Heinrichs[1,2] · **Sebastian Knell**[2]

Received: 19 March 2021 / Accepted: 20 August 2021 / Published online: 23 October 2021

Abstract
In this paper, we use some elements of the philosophical theories of Wilfrid Sellars and Robert Brandom for examining the interactions between humans and machines. In particular, we adopt the concept of the space of reasons for analyzing the status of artificial intelligent agents (AIAs). One could argue that AIAs, like the widely used recommendation systems, have already entered the space of reasons, since they seem to make knowledge claims that we use as premises for further claims. This, in turn, can lead to a sense of alienation because AIAs do not quite play by the rules of the space of the reason. We, therefore, ask somewhat pointedly whether aliens have entered the space of reasons. A closer look reveals that it is a misconception to consider AIAs as being (already) in the space of reasons. In fact, they should be seen as very sophisticated tools. Since these tools affect our own acting in the space of reasons, special regulations are required for their proper use.

Keywords Human–machine interaction · Artificial intelligent agents · Space of reasons · Responsibility · Ethics

1 Introduction

In a famous and often-quoted passage of his 1956 essay *Empiricism and the Philosophy of Mind*, Wilfrid Sellars claimed:

This article is part of the Topical Collection on *Information in Interactions between Humans and Machines*

✉ Bert Heinrichs
b.heinrichs@fz-juelich.de

[1] Institute of Neurosciences and Medicine: Ethics in the Neurosciences (INM-8), Forschungszentrum Jülich, 52425 Jülich, Germany

[2] Institute of Science and Ethics (IWE), Rheinische Friedrich-Wilhelms-Universität Bonn, Bonner Talweg 57, 53113 Bonn, Germany

> 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 (1997, § 36).

From then on, the metaphor "space of reasons" took its rise in philosophy. It became even more important after Robert Brandom picked it up and gave it further theoretical interpretation. Today, it is widely used for designating the inferentially structured sphere of propositional mental states and verbal utterances. By its logical structure, the space of reasons differs fundamentally from the causally structured "realm of law" (McDowell, 1994, 97).

According to Brandom, the space of reasons is closely connected with a number of other core philosophical concepts. Especially important is the fact that the space of reasons is socially structured. In a paper from 1995, Brandom explicated:

> 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 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 (1995, 903-904; see also Brandom, 1994, 199-206).

At first glance, these rather abstract considerations may not seem like the best starting point for thinking about human–machine interaction, especially if one is primarily interested in ethical and social aspects of this interaction. In contrast, we think that Sellars and Brandom provide an especially rich philosophical concept that makes it possible to explore human–machine interaction and that allows for working out ethical implications in particular. In this paper, we want to do just that. To be sure, we are not concerned with a philosophical examination of the space of reasons itself or with an interpretation of Sellars' or Brandom's philosophy. Rather, we use the notion of the space of reasons to assess issues in human–machine interaction that we think are especially important. To this end, we will first analyze the above quotation from Brandon in detail in order to prepare the basis for our further considerations (II). Subsequently, we will examine how artificial intelligent agents (AIA) appear to have already entered the space of reasons (III). Then, we will argue that their position in everyday life (or our attitude towards them) can cause a

certain form of alienation among the established inhabitants of the space of reasons, namely humans (IV). Finally, we will conclude that it is a misconception to consider AIAs as being (already) in the space of reasons. Rather, they should be seen as very sophisticated tools (V)—tools, however, which affect our own acting in the space of reasons, and which, in turn, require special regulations for their proper use (VI).

2 The Social Articulation of the Space of Reasons

Let us start by looking at the above quote a little more closely. One of the core ideas of Brandom's approach—notable one that can be traced back to Plato and that Brandom shares with many contemporary philosophers—is that to know something (or, more generally, to achieve the status of a knower) means to be able to provide reasons for it. And this, in turn, is a social act between (at least two) persons. As Sellars pointed out more than half a century ago, if we characterize a state as knowing, we are not providing an empirical description. Rather, we are maintaining—to use the Brandomian phrase—that someone is taking part in the game of giving and asking for reasons. Moreover, in this game, to claim to know something entails (1) undertaking a commitment (oneself), (2) attributing an entitlement (to others) to ascribe that commitment, and (3) attributing a further entitlement (to others), namely the entitlement to undertake the same commitment himself or herself and to refer to the first speaker in case the entitlement is put in question by a third person. Additionally, one can conceive of the first speaker as undertaking a second-order commitment to give reasons that demonstrate his or her entitlement to the primary commitment in case such reasons are asked for.[1] Or, to put it another way, in order to assume knowledge, it is not enough for an agent to have a true and justified belief—he or she must be able to perform certain moves in the space of reasons.

This is best illustrated by a simple example: Imagine Alice says, "It will rain this afternoon." ("I know that it will rain this afternoon.") She, then, commits herself to providing reasons if asked for. Suppose Bob is unconvinced and asks Alice, "How do you know?" ("Can you provide a reason for your claim?") Alice might answer, "I saw the weather forecast this morning." Bob might continue to be skeptical and ask again, "Are you sure? The weather looks fine to me and the weather forecast is sometimes wrong." Alice could reply, "Yes, that's true. But today, they showed a satellite picture indicating that rain clouds are moving fast in our direction. I am sure it will rain." By making the initial claim, Alice is undertaking a commitment to provide reasons and, vice versa, Bob is attributing this commitment to Alice. At the same time, Alice is attributing an entitlement to Bob namely, to use the claim, "It will rain this afternoon." as a premise in further inferences. Bob could, for example,

[1] It is a somewhat tricky thing to determine, whether this second-order commitment really is an analytically separable and additive commitment or whether it must rather be conceived as being already implicitly entailed in the primary commitment, thereby contributing to its full normative force. Brandom himself is speaking of a "task responsibility" with respect to this reason giving procedure (Brandom, 1994, 173).

say to Carol, "We should cancel the BBQ for this afternoon." If asked by Carol why he thinks so, Bob is legitimate to say, "Alice told me that it will rain." The example also highlights an additional feature of the social structure of the game of giving and asking for reasons: One can use an interlocutor's commitments in one's own collateral commitments, notably in those which are not necessarily shared by the interlocutor. By linking the commitment concerning the rainy weather with the practical commitment to give a BBQ in the afternoon one can draw the conclusion that the latter practical commitment is incompatible with the former assertional one and thus has to be dropped.[2]

All of this may seem pretty trivial. It is not. Actually, Brandom unfolds his inferentialism to a full-blown theory of meaning and intentionality, with the specific representational dimension of the latter being explained in terms of the social perspectival articulation of collateral commitments (Brandom, 1994, Ch. 8). However, we do not have to follow this long and arduous path here any further. It is enough to be aware of the importance of the space of reasons, the game of giving and asking, and, most importantly, the social structure of both of them.

3 Artificial Intelligent Agents

Artificial intelligent agents (AIAs) are already omnipresent today and significantly shape our daily lives. Following an established usage of the term, by AIAs, we refer to various forms of computer systems "that can decide what to do and do it." (Russell & Norvig, 1995, viii). In particular, we focus on the system which incorporates deep learning algorithms. In their seminal 2015 paper, AI pioneers Yann LeCun, Yoshua Bengio, and Geoffrey Hinton explain the term "deep learning" as follows:

> Representation learning is a set of methods that allows a machine to be fed with raw data and to automatically discover the representations needed for detection or classification. Deep-learning methods are representation-learning methods with multiple levels of representation, obtained by composing simple but non-linear modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level. [...] The key aspect of deep learning is that these layers of features are not designed by human engineers: they are learned from data using a general-purpose learning procedure. (LeCun/Bengio/Hinton, 436).

The most widespread methods of deep learning at present are supervised learning, unsupervised learning, and reinforcement learning. Without going into the technical details, the crucial point in each case is that there are deep layers which "can be seen as distorting the input in a non-linear way so that categories become linearly separable by the last layer" (LeCun/Bengio/Hinton, 438). Note that the categories are not predefined by humans but are formed and refined independently by

[2] Brandom analyzes this social perspectival form of drawing inferences in detail in Brandom 1994, chap. 8.

the algorithms. In this sense, they are "learning" systems. In philosophy, there are very different and sometimes highly sophisticated concepts of learning. For example, the term "learning" can be closely related to concept use and rationality. If one does this, one obtains a form of learning, which must take place in the space of reasons. For the process of learning is then mediated by the adoption of new—and better—reasons. In contrast, Ludwig Wittgenstein uses a more basic notion of learning, namely learning as a mere form of behavioral training by confrontation with examples (Wittgenstein, 2009 §§ 5–6). Deep learning procedures obviously resemble this basic type of learning in some aspects. In the context of this work, it should be clear that we do not consider the learning procedures of deep learning algorithms as already taking place in the space of reasons. Otherwise, it would not be an open question anymore whether these systems can inhabit this space.

We do not want to limit ourselves to one particular method or type of algorithm here. The field is developing dynamically, and there is now a whole range of hybrid approaches. We also do not want to use the term "agent" in any specific way, but rather as a general term to describe artificial systems that can have an impact on their environment that fulfills certain additional conditions. Having an impact on the environment alone is certainly not enough. After all, this obviously applies to natural events as well, which we hardly want to call "agent." Luciano Floridi has suggested "interactivity," "autonomy," and "adaptability" as further criteria (Floridi, 2013, 140). Our inclusive understanding of AIA is inspired by Floridi's approach of widening the notion of an agent.

Typical examples of AIAs in the sense we have in mind and we are dealing with include recommendation systems such as Amazon's, Netflix's of Spotify's, or personal assistants such as Apple's Siri, Amazon's Alexa, or Google's Assistant. As noted, they are widespread and affect our lives deeply. The type of influence such systems have has long since ceased to be incidental. Even though one need not be an Amazon customer or have a Facebook account, of course, a complete refusal to use services of the kind mentioned is only possible today if one is willing to refrain from participating in large parts of public life. Many would say that they make life easier. Others are skeptical or even negative about them. What we are interested in here is that they seem to make knowledge claims.

Take a typical example: If you buy a book at Amazon's you get a recommendation for other books which could be of interest for you. The more products you buy, the more accurate the recommendations usually become. The AIA in the background learns from previous purchases and can thus refine its recommendations for you (Smith & Linden, 2017; for an interesting analysis of Spotify's recommendation engine cf. Huq, 2019). In a way, the AIA appears to make a knowledge claim when suggesting, "You will like this product." ("I know that you will like this product.") What is more, this knowledge claim often seems to be well justified. Even if many do not like to admit it, the recommendations regularly hit the mark. The AIA really seems *to know* what we like or what we are interested in. And even if a recommendation proves to be wrong, it still has the logical form of a knowledge claim. In fact, only because it does have this logical form, it can be wrong in the first place.

Take another example: AIAs playing board games. Since the early days of AI, board games were a favored playfield for researchers. One event that attracted

worldwide attention was the victory of IBM's chess program Deep Blue against the then world champion Gary Kasparov in 1997. It is interesting that Deep Blue's chief developer, Feng-Hsiung Hsu, does not classify the program as AI. In fact, he describes the skepticism and even rejection that existed among some of his collogues about AI (Hsu, 2004). Accordingly, he regards Deep Blue as a mere tool and consequently the match against Kasparov not as a clash of man and machine, but rather as man-as-player against man-as-toolmaker. Actually, there are good reasons for this view, which have to do with the way Deep Blue works. Anyway, the situation has fundamentally changed with AlphaZero. This is a Go program that was not trained using saved Go games, but only by playing games against itself. Within a short time, it was much stronger than a previous version, which won against Go pro Lee Sedol in 2016. The developers of AlphaZero maintain:

> Humankind has accumulated Go knowledge from millions of games played over thousands of years, collectively distilled into patterns, proverbs and books. In the space of a few days, starting tabula rasa, AlphaGo Zero was able to rediscover much of this Go knowledge, as well as novel strategies that provide new insights into the oldest of games (Silver et al., 2017, 358).

It seems appropriate to say that AlphaZero makes a knowledge claim through its innovative game play quite comparable to a human player who implicitly asserts, "This is the best move."

Take a final example from the field of medicine. In 2017, Andre Esteva and colleagues published a paper in *Nature* in which they presented a convolutional neural network (CNN) that was trained with a dataset of 129,450 clinical images for the classification of skin lesions (Esteva et al. 2017). Subsequently, it was tested against 21 board-certified dermatologists on biopsy-proven clinical images and achieved a performance on par with all tested experts. At the end of the paper, the authors suggest that mobile devices equipped with the software could extend the reach of dermatologists outside of the clinic and potentially provide low-cost universal access to vital diagnostic care. Suppose the AIA would be available already and you would have it on your mobile phone. Every now and then, you could use your phone's camera to examine suspicious skin areas. The system would then, hopefully, report, "There is no skin lesion of concern." and thereby apparently make a knowledge claim.

Considering these three examples, one can very well get the impression that we are no longer alone in the space of reasons. The moment AIAs began to express themselves in terms of knowledge claims, they seemingly entered the space of reasons. At least in one respect, we seem to have long since accepted them in our epistemological neighborhood: We use their statements without further ado as premises in our own claims. Or would you find it strange if someone would answer, "Amazon recommended it to me." to the question "How do you know you will like this book?" Probably not (unless you belong to the skeptics mentioned above). In other words, we feel *entitled* to a claim something based on a prior claim by an AIA. This,

apparently, indicates that AIAs are players in the game of giving and asking for reasons in their own right. However, this leads to a serious problem. Before we go into it in more detail, let us look at certain kind of discomfort that goes along with the widespread use of AIAs.

4 Alienation in the Space of Reasons

We already mentioned above that some people feel uncomfortable with AIAs. The reasons for this are certainly manifold. Sometimes, a broader skepticism of technology or even hostility to technology lurks in the background, sometimes it is simply the fear that cherished practices are in danger of disappearing as AIAs become more widespread. For example, some regret that small bookstores where they received individual recommendations are being squeezed out by large internet companies. Not only do the recommendations from Amazon & Co seem less accurate to them, but they also miss the personal contact and the opportunity to talk to the bookseller about the last book they read. Similar displacement mechanisms are at work in many places. While AIAs undoubtedly have advantages, something is also lost with them. One or the other may already experience this form of change as a form of alienation. In a sense, he or she no longer feels at home in the impersonal world of AIAs. But that alone would hardly be enough to speak of alienation in a deeper sense. It would merely be the familiar pattern that innovations eventually become alienating to people once they get older.

However, it may be possible to identify another form of alienation related to AIAs that reaches deeper and is more fundamental in nature. Apparently, AIAs do not behave quite right in the space of reasons and partly disregard the established rules of the game of giving and asking for reasons. Think about the bookseller again: If she gives you a recommendation, maybe you would ask for more details. Or after a reading, perhaps you would tell her your impressions and discuss the book with her. This is not possible with an AI recommendation system—and this feels weird. When someone tells you something, you expect to be able to ask questions and make comments.[3] If this is not possible, it is a profound deviation from our common discursive practice. The deviation is so serious because AIAs seem to get in our way on our very own territory and seem to set up new rules or at least partially override the old rules. To use the Brandomian terminology, they are not undertaking justification commitments—as they should do when making knowledge claims.[4] In summary and somewhat loosely formulated, one could say: aliens have entered the space of reasons and this leads to a feeling of alienation among the traditional inhabitants.

[3] In fact, Amazon offers a kind of inquiry option with its "Why recommended?" function. Other recommendation systems have similar functions that are there to improve the system through user feedback.

[4] To be more accurate, they are neither committing themselves to a specific knowledge claim p which includes, for example, not to claim non-p, nor are they committing themselves to provide reasons for p. The latter is the more obvious and problematic point.

5 Responsible Agents and Sophisticated Tools

There is a simple reply to this line of argumentation: AIAs are just not "in" the space of reasons. They only seemingly participate in the game of giving and asking for reasons. In fact, however, they are only sophisticated tools which we use just as we use a hammer to drive a nail into the wall. To be clear, we think that this reply is appropriate—at least in part. Yet, it should be taken into account that AIAs do something that hammers do not: they produce knowledge. Again, one could reply that other tools also produce knowledge or, to be more accurate, help us to produce knowledge. Without a thermometer, for example, it is difficult for us to give the exact temperature. Only with its help can we determine, "The temperature is 21.8 degrees Celsius." As a shortcut, we then sometimes say, "The thermometer says it is 21.8 degrees Celsius." In such cases, too, we seem to impute a knowledge claim to the thermometer. However, this is such obvious nonsense that it does not come to our thinking that thermometers might have entered the space of reasons. After all, a thermometer is just a tool and "the thermometer says" is only a colloquial phrase that probably no one takes too serious. There must be a substantial difference between thermometers and AIAs. If this were not so, the feeling of alienation described above would simply not exist or, if it did exist, it would simply be out of place. Yet, we think it is at least partly adequate.

We suggest that it is the extremely high level of complexity that prevents us from thinking of AIAs simply as tools that support us in our actions (which includes producing knowledge). Rather, they give us a perspective on the world that is, at least in part, essentially different from our own.[5] They recognize patterns where we are unable to recognize any. We tend to understand this as the making of genuine knowledge claims. For we do make use of these claims in the context of our own collateral commitments in the social perspectival way described above. I might, for example, wonder that if a system recommends a book to me, my friend might also like it and therefore I will buy it for him or her—thereby making inferential use of the systems claim-like output in the collateral context of my own collateral commitment that my friend has similar preferences concerning literature as me.[6] The widespread feedback functions reinforce this impression by enabling (or at least simulating) a kind of exchange of reasons. On the other hand, however, AIAs are not full-fledged actors in the space of reasons, because they lack a crucial ability: they cannot undertake commitments. As a consequence, they cannot really make knowledge claims,

[5] As we said, we are thinking of deep learning here. For other types of AI, our considerations may not apply or only to a limited extent.

[6] If one follows Brandom's systematic approach a little bit further, then one could even conclude, that by making use of an AIAs utterance as a premise within the context of one's own collateral commitments one implicitly takes the AIA to be a real representer. For according to Brandom's analysis the representational dimension of assertional expressions is made explicit by de-re-ascriptions of their content which, in turn, make explicit the social-perspectival constellation of inferentially exploring that content in the context of one's own collateral commitments. Cf. Brandom 1994, Ch. 8.

for this would entail their undertaking an assertional commitment in a full and unrestricted sense. The challenge is to come to terms with this paradox.[7]

If we think back to Brandom's quote, the key point is that making a commitment involves a normative act constituted by a normative attitude and a bundle of further and consequential normative attitudes—and AIAs seem not to be in a position to do exactly that.[8] They seem unable to understand themselves as standing in genuine normative relationships with other entities and their performances. Note that this is not a dogmatic claim. Perhaps at some point in the future, AIAs can take on the kind of commitments that are constitutive for the game of giving and asking for reasons. This is undecided at present. Currently, they most likely cannot.

But how would we know that they can? AI researchers and many others frequently refer to the Turing test in this context (Turing, 2013). Critics object that being able to fake a conversation is not the kind of capacity needed for entering the space of reasons.[9] Frankly, today no one may know exactly which capacity or set of capacities is required instead. Some philosophers will probably refer to self-consciousness, others will cite reflective cognitive processes or phenomenal consciousness, and yet others will take a completely different view. How one or the other can be empirically proven is again another question. More research is needed to see more clearly here. Finally, there are opinions that having the status of a knower is not something that can be described in purely descriptive terms. At least that is what Wilfrid Sellars was convinced of:

> Now the idea that epistemic facts can be analysed without remainder—even 'in principle'—into non-epistemic facts, whether phenomenal or behavioural,

[7] A similar problem can occur when dealing with higher animals. Let us take the case of a dog barking when a stranger approaches the house, long before a human inhabitant of the house notices this. We could say that the dog knows that a stranger is coming, and we would probably even use this "knowledge" as a premise for further inferences ("I should lock the door, someone is coming!") On the other hand, the dog is not really "in" the space of reasons, because it cannot provide any (linguistically articulated) reasons for its assertion ("A stranger is coming."). This is completely obvious to us, which is not least made clear by the fact that we would not hold the dog accountable in case of a false alarm. (Well, some dog owners would probably say to their dog, "Why did you bark again, although nobody came? I got up to lock the door for nothing." But surely, they should realize that this cannot be entirely serious). In short, higher animals seem to inhabit a borderland of the space of reasons, which is sometimes puzzling to us and which certainly shapes our interactions with them. In contrast to AIAs, however, we cannot treat higher animals simply as tools, because they are living beings that have sensations like we do. It is noteworthy that in the long history of human-animal relationships, this latter fact has not always been considered conclusive. In fact, animals have often been regarded as mere tools or "soulless machines.".

[8] Also, the dog mentioned above does not seem to be able to understand and accept this elaborated type of normative attitude.

[9] Brandom has expressed the view that the Turing test is quite appropriate Brandom, 2008(, 69–77). Maybe his rather liberal attitude with respect to this question has to do with a kind of residual behaviorism that his general theory of discursive practice entails, according to which normative attitudes can just be implicit in sanctioning behavior. Cf. Brandom, 1994, Ch. 1. However, we are skeptical about this reductionist claim. In particular, it is important to note that having a conversation is different from exchanging reasons. Conversations often consist to a large extent of statements that are not knowledge claims, but rather questions and expressions ("How are you?," "Oh, thank you, good. And you?") or which, while ostensibly knowledge claims, are in fact purely polite phrases that are not designed to be substantiated in more detail ("Your lecture was fantastic!").

> public or private, with no matter how lavish a sprinkling of subjunctives and hypotheticals is, I believe, a radical mistake—a mistake of a piece with the so-called 'naturalistic fallacy' in ethics (1997, § 5; see also 2007, 406-408).

We should keep this in mind when thinking about AIAs and our relationship with them. Moreover, we should consider the possibility that the space of reasons does not have sharp boundaries but has gray areas at its edges (to speak very metaphorically) or that there are inhabitants of qualitatively different kinds.

6 Living with AIA

For the time being, we probably should consider AIAs merely as complex tools and not as agents in the space of reasons (Heinrichs & Eickhoff, 2020). However, even if we do so, a problem remains regarding their proper use. Remember that making a knowledge claim entails being able to provide reasons for that claim when asked. Now imagine you are a doctor who uses an AIA to make medical diagnoses. Based on a variety of medical data, the system indicates for one of your patients that he or she is affected by a disease that was clinically inconspicuous until now. Of course, you should tell your patient the result that the AIA generated and suggest a therapy if available. However, since the AIA is only a tool, the result of the AIA alone cannot be considered a knowledge claim. In particular, the AIA does not authorize and thereby entitle you to make a diagnosis. Rather, it provides only a piece of evidence, such as the temperature indicated by a clinical thermometer or the heart and lung sounds, that can be heard with the aid of a stethoscope. If the patient asks for reasons for the diagnosis, you can cite all of those. But unlike the clinical thermometer and the stethoscope, the AIA is so complex that you cannot check its output yourself. What is more, the complexity is such that hardly anyone or even no one can test its validity (apart, of course, from the statistical evidence that results from multiple applications in the past). That is one reason why it seems to us that the AIA itself makes a knowledge claim in the first place. But if we must blindly rely on the system, then perhaps we had better not consider it as a source of evidence at all. Epistemic standards suggest not relying on evidence that is not comprehensible to us. As with all tools, the users are responsible for the final output—not the tool. But if the tool is totally opaque, no user can take this responsibility. He or she would not be able to fulfill his commitment namely, to provide reasons for knowledge claims.

Under the title "explainable AI," there are intensive efforts to solve exactly this problem (Samek et al., 2019). The goal is, roughly speaking, to increase the epistemic transparency of AIAs so that users can better assess why an output was generated. This should enable users to decide what role a result can play in the context of a comprehensive judgment.

What does all this mean for human–machine interaction? First, there are no new inhabitants in the space of reasons so far. Due to their enormous complexity, it may sometimes seem as if AIAs are making knowledge claims. In fact, however, they lack the ability to undertake commitments of justification. Furthermore, they seem even unable to adopt the genuine normative attitude of undertaking a primary assertional

commitment at all. It follows, secondly, that even very complex AIAs are only tools. This could change at some point in the future, and it remains unclear for the time being how we will know when the situation has changed.[10] Thirdly, it means that we must design AIAs so that we can use them responsibly as tools, i.e., without running the risk of not being able to meet commitments on our part. Eventually, there can be no aliens in the space of reasons. If someone is really at home in the space of reasons, then he or she cannot be a complete stranger to us. Inhabitants of the space of reasons, however, can behave strangely, namely when they do not fulfill their discursive obligations. This last warning is addressed to us, not to AIAs. For now, we alone bear the responsibility for the ethical design of human–machine interactions and this includes, among other things, giving reasons for how we refer to results provided by AIAs.

In closing, we may take a thought further, put forward by Brandom in the opening passage of the first chapter of *Making it Explicit*: "We are the ones for whom reasons are binding, who are subject to the peculiar force of the better reason." (Brandom, 1994, 5) 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 AIAs. But it says something very fundamental about those who use it—it characterizes them as *normative*. Questions concerning human–machine interactions are, eventually, questions about us as normative beings. One day, it may be that AIAs have attained the status of normative entities adopting genuine normative attitudes essential for speech and concept use. Then, they would belong to "us." While we are actually talking about human-tool interactions today, we would then have to talk about how to deal with other normative subjects.

Author Contribution Both authors contributed equally to the paper.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability n. a.

Code Availability n.a.

Declarations

Conflicts of Interest The authors declare no competing interests.

[10] An important criterion, of course, would be that we be able to attribute to AIAs genuine normative attitudes of the complex kind that Brandom describes as constitutive of undertaking and keeping score on commitments. But it is not a straightforward matter to answer the question of exactly what kind of behavioral evidence might warrant such normative attribution. As we noted above (Note 10), we are rather skeptical of Brandom's own approach to this question.

References

Brandom, R. B. (1994). *Making it explicit. Reasoning, representing, and discursive commitment*. Harvard University Press.

Brandom, R. B. (1995). Knowledge and the social articulation of the space of reasons. *Philosophy and Phenomenological Research, 55*(4), 895–908. https://doi.org/10.2307/2108339

Brandom, R. B. (2008). *Between saying and doing*. Oxford University Press.

Esteva, A., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*, 115–118. https://doi.org/10.1038/nature21056

Floridi, L. (2013). *The ethics of information*. Oxford University Press.

Heinrichs, B., & Eickhoff, S. B. (2020). Your evidence? Machine learning algorithms for medical diagnosis and prediction. *Human Brain Mapping, 41*, 1435–1444. https://doi.org/10.1002/hbm.24886

Huq, P. (2019). Music to my ears: De-blackboxing Spotify's recommendation engine. https://blogs.commons.georgetown.edu/cctp-607-spring2019/author/ph625/. Accessed 10.09.2021.

Hsu, F.-H. (2004). *Behind Deep Blue: Building the computer that defeated the world chess champion*. Princeton University Press.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*, 436–444. https://doi.org/10.1038/nature14539

McDowell, J. (1994). *Mind and world*. Harvard University Press.

Russell, S., & Norvig, P. (1995). *Artificial intelligence: A modern approach*. Pearson.

Sellars, W. (1997). *Empiricism and the philosophy of mind*. Harvard University Press.

Sellars, W. (2007). Philosophy and the scientific image of man. In K. Scharp & R. B. Brandom (Eds.), *In the space of reasons. Selected essays by Wilfrid Sellars* (pp. 369–408). Havard University Press.

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (Eds.). (2019). *Explainable AI: Interpreting, explaining and visualizing deep learning*. Springer.

Smith, B., & Linden, G. (2017). Two decades of recommender systems at Amazon.com. *IEEE Internet Computing, 21*(3), 12–18. https://doi.org/10.1109/MIC.2017.72

Silver, D., Schrittwieser, J., Simonyan, K., et al. (2017). Mastering the game of Go without human knowledge. *Nature, 550*, 354–359.

Turing, A. M. (2013). Computing machinery and intelligence. In B. J. Copeland (Ed.), *The essential Turing* (pp. 441–464). Clarendon Press.

Wittgenstein, L. (2009). *Philosophical investigations*. 4th ed. by P.M.S. Hacker and Joachim Schulte. Wiley-Blackwell.