This is a preprint. Citations should be to the published version in Hakli, Mäkelä & Seibt (eds) Social Robots in Social Institutions. (2023) IOS Press.

doi:10.3233/FAIA220628

The AI-Stance: Crossing the Terra Incognita of Human-Machine Interactions?

Anna STRASSER[a,1] and Michael WILBY[b]
[a] *LMU, Munich, Germany*
[b] *Anglia Ruskin University, Cambridge, UK*

Abstract. Although even very advanced artificial systems do not meet the demanding conditions which are required for humans to be a proper participant in a social interaction, we argue that not all human-machine interactions (HMIs) can appropriately be reduced to mere tool-use. By criticizing the far too demanding conditions of standard construals of intentional agency we suggest a minimal approach that ascribes minimal agency to some artificial systems resulting in the proposal of taking minimal joint actions as a case of a social HMI. Analyzing such HMIs, we utilize Dennett's stance epistemology, and argue that taking either an intentional stance or design stance can be misleading for several reasons, and instead propose to introduce a new stance that is able to capture social HMIs – the AI-stance.

Keywords. human-machine interactions, social agency, stance epistemology, distributed responsibility

1. Introduction

It is likely that we will soon share a large part of our social lives with various new kinds of interactive artificial systems, and it is therefore, at least, conceivable that we might sooner or later consider them as social agents instead of mere tools [1, 2]. Even if we do not attribute full-blown mentality to artificial systems, it seems at least plausible that we may soon naturally respond to interactions with artificial systems as we do in social interactions with humans. Indeed, there are already human-machine interactions (HMIs) that cannot satisfyingly be reduced to mere tool-use [3], and this is because – unlike mere tools – artificial systems based on learning algorithms, such as some social robots, are able to act with some degree of autonomy, to learn from experience (and adapt their goals correspondingly), and to react to social cues. Our aim in this paper is, first, to elaborate to what extent certain HMIs could be regarded as a new type of social interaction – different in kind to those which we might engage in with other adult humans, children, or non-human animals – and then, second, to investigate what consequences this has for how to explain social HMIs. Utilizing Dennett's stance epistemology, we shall argue that taking either an intentional stance or design stance can be misleading for several reasons, and instead propose to introduce a new stance that is able to capture HMIs that cannot be reduced to tool-use. In short, we will argue that the adoption of an intentional stance is too easily associated with far too demanding conditions regarding the agency of participants and opens the door for implicitly but inappropriately ascribing full-fledged

[1] Corresponding Author, Anna Strasser; E-mail: annakatharinastrasser@gmail.com.

moral agency. Consequently, we suggest considering a more sophisticated and subtle stance – the AI-stance.[2]

Standard philosophical conceptions describing human-human interactions are based on a dichotomy between action and behavior, separating intentional actions from mere behavior.[3] It is often presupposed that machines, in principle, are not able to act intentionally; whatever they do is just described as 'mere behavior'. Consequently, all interactions with machines are described as *tool-use* and not as *social interactions*. However, if there are HMIs in which artificial systems do not just behave – as mere tools do – then we should not categorize such interactions as mere tool-use. Acknowledging that such HMIs nevertheless do not fulfill the demanding conditions of prototypical social interactions as we find them in human-human interactions, we find ourselves in a *terra incognita* for which we have no established notions yet. To capture such interesting in-between phenomena, we have to overcome standard construals of agency and sociality that suggest that social agency only applies to conceptually sophisticated living beings. We will do this by taking joint actions as one paradigmatical example of social interactions.

2. Two Theories of Joint Action Rejected

What does it take to be a fully-fledged social agent with the capacity to engage in joint action? We will consider, and then reject, two popular answers to this question: first, *an intellectualist conception* of intentionality, and second, a *biological conception* of intentionality. Both accounts lead to the conclusion that artificial systems are in principle not capable of engaging in joint action. Our aim in this section is to explain why these popular accounts should be rejected, which will then leave us with the default assumption that some artificial systems are capable of participating in joint action.

2.1. Intellectualist Conceptions of Intentional Action

On a rich, intellectualist conception of intentional agency, it is claimed that one needs to be in possession of a complex suite of conceptual resources to be an intentional agent. This, for instance, is the position of Donald Davidson [6-11], who argues that constitutive relations holding between propositional attitudes and their contents, as well as language, intentional action, and interpretation, sharply separate off 'the beasts' from rational animals such as humans: "the intrinsically holistic character of the propositional attitudes makes the distinction between having any and having none dramatic" [9]. We can assume that the current limitations of all artificial systems – with regards to the lack of flexibility and systematicity in artificial thought – would put all 'automata' on the wrong side of Davidson's 'dramatic' divide.

The arguments that Davidson provides in favor of this are complex and varied.[4] We will briefly review one major strand of the Davidsonian argument as sketched in this passage: "To have a single propositional attitude is to have a largely correct logic, in the

[2] The abbreviation 'AI' should be understood as a placeholder not just for artificial *intelligence*, but more specifically for artificial *intentionality*.

[3] This distinction can be made using a variety of different terms. Sometimes the distinction is drawn between *action* and *bodily movement* (e.g., [4, 5]).

[4] The most careful comprehensive overview of Davidson's work as a whole is [12], and its more technical companion [13].

sense of having a pattern of beliefs that logically cohere. This is one reason why to have propositional attitudes is to be a rational creature. The point extends to intentional action. Intentional action is action that can be explained in terms of beliefs and desires whose propositional contents rationalize the action" (Davidson [9], p. 99).

The central claims are, first, that intentional action requires belief/desire pairs, and second, that belief/desire pairs cannot exist as isolated, atomic elements of an entity's cognitive economy, but must come as part of a logically coherent holistic pattern of propositional attitudes. The main argument in favor of the first claim is that one can designate an agent as the *author* of an action only insofar as that action is *ascribed* to the agent. To ascribe an action to an agent is to understand that action as 'intentional under a description', where such a description recounts the agent's reasons for the action, and where the reasons are a belief/desire pair that both *rationalize* and *cause* the action.

With regards to the second claim, there are various interrelated arguments for the viewpoint that come with a good degree of background theoretical commitments which we do not necessarily share, so we can focus on one particular argument that has some independent force. This is the idea that it doesn't make sense to attribute a singular belief/desire pair to an agent. Take, for instance, the belief/desire pair that comprises the intention to pour a cup of coffee. This intention would lack content if it were to be assumed that the agent didn't understand what it was to 'pour' something; that liquid is something that you can pour; etc. Take away one of these beliefs – e.g., the belief that coffee (in this case) is a liquid – and the original belief/desire pair that we attribute to the agent no longer rationalizes their action, and so, no longer makes it intentional under that description.

Together, then, these two claims provide an account of intentional agency that requires of any intentional action that it be carried out by an entity with an integrated, holistic set of propositional attitudes.

Davidson's intellectualist approach perhaps tells us something interesting about the end-point of a human maturation process, where capacities for thought, language and interaction intersect in deeply integrated and holistic ways. However, there is good reason to suppose that it cannot be the whole story about being a participant in social interaction in general. There are two families of objections to the Davidsonian view: empirical-based objections, and conceptual-based objections.

Empirical-based objections point to apparent counterexamples in the developmental and comparative psychology literature. For instance, evidence suggests that there are multiple realizations of socio-cognitive abilities in various types of agents such as infants and non-human animals [14-19]. Such subjects have certain socio-cognitive abilities which strongly suggest that they can be active participants in joint actions. This evidence supports the idea that it is not only conceptually sophisticated humans with whom one can interact socially, so also suggests that it might likewise become part of our common sense to consider certain artificial systems as social interaction partners.

Conceptual-based objections object to the sharp 'all-or-nothing' dramatic divide that the Davidsonian thesis inserts between entities. This sharp divide all but rules out the idea that there is a gradual, learnable maturation process that allows for a transition between those entities that are only weakly intentional and those entities that are richly intentional. But this is implausible. It would require a sudden saltation from a non-intentional entity to an entity with a fully-formed integrated system of propositional attitudes to occur both at the phylogenetic level (in the shift between non-human animals to human animals) and at the ontogenetic level (in the shift from infanthood to adulthood). There is reason to think that in both cases, the shift is much more gradual than this and

is at least partly about learnable procedures (see for instance [20-21] for the ontogenetic case, and [22-23] for the phylogenetic case).

2.2. Biological Conceptions of Intentional Agency

In contrast to rich intellectualist accounts such as Davidson's, one might instead argue that any kind of agency that enables entities to be a participant of a joint action requires internal affective states (emotional, mental, and conscious states). For instance, John Searle concludes his famous discussion of the limitations of 'strong AI' by saying: "mental states are biological phenomena. Consciousness, intentionality, subjectivity, and mental causation are all a part of our biological life history, along with growth, reproduction, the secretion of bile, and digestion" [24, p. 41]. For Searle, AI is not capable of engaging in joint action because AI lacks the biological make-up to have genuine intentional and conscious thoughts. Such a view assumes, therefore, that entities that lack consciousness, mental and emotional states can only behave – not act – and therefore it is concluded that artificial systems cannot qualify as social interaction partners and, consequently, every human-machine interaction should be understood as mere tool-use.

Instead of disqualifying machines because they are not living, biological beings, however, one can instead consider starting from the assumption that the way living beings fulfill the conditions for agency is just *one way* to realize agency. Assuming that there are multiple realizations of agency possible, one can extend the conception of agency in various interesting ways [25-27].

To conclude this section, it is useful to contrast why infants, on the one hand, and artificial systems, on the other, are thought by some to fall short of genuine social agency. With infants, the supposed shortfall comes with the capacity to recognize genuine mentality in others. Notoriously, young children fail the (explicit) false-belief task and related tests in understanding others [28]. With artificial systems, the supposed shortfall is with their own lack of mentality: biological constraints are thought to exclude them from agency right from the start. We find neither constraint useful since neither constraint is, as we have argued, necessitated on either a priori or empirical grounds. At this stage of the development of AI it is, we think, better to keep an open mind about the nature of HMI and to explore other avenues for how social HMI might be possible.

3. Towards gradual approaches

We have argued that, although two of the standard approaches to intentional agency end up denying that current artificial systems are capable of engaging in joint intentional agency, these arguments lack conviction. We see no reason to suppose that HMIs should necessarily be designated as sub-intentional, especially when those interactions strike the human contributor intuitively as cases of genuine shared agency.

The idea that there is a mid-way point between rich, intellectualist views of shared agency on the one hand, and sub-intentional interactions that amount to 'mere behavior' on the other, has been explored in other areas in recent years, especially under the label 'minimalism' [29]. Positions along these lines have been adopted by thinkers such as Elisabeth Pacherie [30], Stephen Butterfill & Ian Apperly [31], John Michael and colleagues [32] and others, all of whom argue that several presuppositions for joint agency can be achieved with cognitive resources that are contentful and representational,

but that needn't be part of a richly interconnected system of propositional attitudes. The main motivation behind these approaches has been to account for the apparent intentional behavior of infants and young children who do not yet satisfy the rich intellectualist demands of a Davidson-style theory, but who clearly do engage in social interactions.

Acknowledging that full-fledged agency might be restricted to sophisticated human beings, we argue that minimal approaches present a promising starting point to expand our conceptual framework to capture interesting in-between phenomena. We argue for the claim that there are HMIs that constitute a stage in-between social human-human interaction and mere tool-use and propose a minimal notion of joint action (illustrating one form of a social interaction) specifying minimal necessary conditions for each type of the involved social interaction partners, thereby, acknowledging an asymmetry between humans and machines (for details see [26-27]).

4. Stance Epistemology

If, as we have argued, HMIs take on an unusual form, then we will need a framework for understanding them. The obvious place to look when thinking about the epistemology of interactions is Daniel Dennett's stance epistemology [33-34]. Dennett has argued that, for any item in the world, we can take one of three 'stances': *The physical stance* is appropriate to explaining physical objects in general; *the design stance* to objects that fulfil a fixed purpose; and *the intentional stance* to objects that operate according to intentional, belief/desire explanations. In this section we shall argue that neither the design stance nor the intentional stance are satisfactory ways of understanding and explaining the new type of HMIs (and we take it as self-evident that the physical stance is uninformative in this respect). As a consequence, we shall recommend that we require an *AI-stance* that differs from the other stances. What is involved in taking the AI-stance, and why it works, shall be outlined in Section 5.

4.1. The Intentional Stance

Following Dennett, one could take an intentional stance to explain those HMIs which cannot satisfyingly be reduced to tool-use. Dennett suggests that the intentional stance can be appropriate to computers (e.g., chess-playing computers) because this stance equips us with successful anticipations of their behavior [33].

However, analyzing the part of artificial systems in HMIs according to the intentional stance can recall the limitations we discussed above, postponing a justification of taking this stance to a probably far-away future in which artificial systems have reached artificial *general* intelligence (AGI). Taking a full-fledged intentional stance is not appropriate to explaining and anticipating the contributions artificial agents with a minimal agency can make in a social HMI.

Another reason – independent of Dennett's epistemological stance terminology – draws on the moral implications which some commentators implicitly assume would follow from taking an intentional stance.[5] According to these commentators, once an agent is capable of acting intentionally, then we should regard it as potentially

[5] Note that this is not a position that Dennett himself takes, since he recommends another stance (what he calls the personal stance [34]) when considering systems as *ethical* beings.

responsible for its actions.[6] But if it is responsible for its actions, it should be regarded as a fully-fledged *moral* agent. Yet, if this is the consequence of taking the intentional stance, it seems ill-fitted to current artificial systems because, intuitively at least, current artificial systems cannot have moral agency [36]. It makes little sense to morally blame one's computer (even a sophisticated one) for, say, breaking down in the middle of an important task.

Acknowledging, therefore, that artificial systems in our society do not have the role of full-fledged moral agents, and acknowledging too that no artificial systems will likely reach a stage of AGI any time soon, we argue that we need an in-between stance to appropriately describe social HMIs.

At this point, we have to clarify that talking about explanations of interaction partners we cautiously distinguish between taking a stance towards a system *in general* (say, as an observer) from taking a stance towards a system as an actual *interaction partner* in a social interaction.[7] This means, in this paper we only argue for the claim that taking an AI-Stance is appropriate with respect to artificial systems which are involved in a social human-machine interaction. Future research has to explore whether there are other incidents, e.g. certain machine-machine interactions, for which taking an AI-stance is also appropriate.

4.2. The Design Stance

Dennett defines the Design Stance as "where one ignores the actual (possibly messy) details of the physical constitution of an object, and, on the assumption that it has a certain design, predicts that it will behave *as it is designed to behave* under various circumstances" ([34], pp. 16-17). It is to be remembered that taking the design stance is very much a matter of adopting an explanatory *stance*, rather than an attempt (either implicitly or explicitly) to uncover the actual intentions of the designer of the artifact; although the latter can help with determining the best stance to take.

Given, as we have argued, that the intentional stance does not adequately capture the role of artificial systems in social HMIs, then perhaps the design stance can. Indeed, Dennett himself suggests that taking the design stance towards computers would allow one to "predict its behavior with great accuracy and reliability" apart from where there is "physical malfunction" ([34], p. 17), although, as noted above, at other points he suggests that the intentional stance can be appropriate to computers (specifically chess-playing computers) [33].

What we shall argue here is that, as with the intentional stance, the design stance does not adequately capture the behavior of certain artificial systems. Whereas the intentional stance *over*-intellectualizes such systems, the design stance *under*-intellectualizes them. For, within the confines of its own design, certain artificial systems act according to intentional and rational patterns. Yet those rational patterns do not get full reign as they would when taking the intentional stance, because they are limited to the domain in which they were designed to be employed. To see this, note that there are two aspects to taking a stance: first you decide to take the stance towards an object, then you work out its behavior by the lights of the procedures suitable to that stance. The

[6] For instance, those theorists who fall into what Behadi and Munthe [35] call the 'functionalist view' of moral agency would appear to subscribe to a view along these lines, drawing a tight connection between intentional agency and moral agency.

[7] For useful discussion of the different ways that stances can be used, see [37].

difficulty with artificial systems is that they present a radical disconnect between these two parts of stance-taking procedure. That is, one cannot, as one would with a human being, take the intentional stance *tout court* to an artificial system. Rather, one is limited only to those domains the artificial system was designed to engage in; outside of that domain, the intentional stance will have no grip.

5. The AI Stance

We have argued that neither the intentional stance nor the design stance present useful ways for understanding the role of artificial systems in social HMIs. In this section we shall argue for a distinctive 'AI-stance' that lies in-between the intentional and design stances. Such a stance allows for a more asymmetric conception of joint action within HMI, and for a corresponding asymmetric conception of moral responsibility within joint actions in which one participant, the artificial system, has only minimal agency. That is to say, it allows for the idea that morally significant joint action can be achieved by agents which share only a minimal commonality in their capacities for intentional and moral thought.

5.1. Means-Ends Reasoning

When one acts intentionally, one usually does so with a goal in mind. We can think of a goal as an outcome – a possible state of affairs that one aims to realize. There are usually various means by which a goal can be realized. In joint action, one can think of the agents as contributing in various ways to a shared goal [38]. That is to say, they share the same token end but might contribute different means towards that end.[8] As Michael Bratman has observed, in order for interactions such as this to be truly collaborative, then it is not just that the interactants need to have meshing means, they also need to both be *intending* the end [40]. For instance, it is not a truly collaborative joint action if two agents are making hollandaise sauce together, but one of the agents is doing their part only under duress, without consent (e.g., at gunpoint from a crazed Gordon Ramsey). Even without putting too rich a set of conditions on what it is to intend an end, we can see that this causes a problem for the idea of HMIs, since it is not clear that any existing artificial system *can* intend an end. As Dennett observes, when discussing Watson (the IBM AI that won at Jeopardy): "It is the absence of practical reason, of intelligence harnessed to pursue diverse and shifting and self-generated ends, that (currently) distinguishes the truly impressive Watson from ordinary sane people. If and when Watson ever reaches the level of sophistication where it can enter fully into the human practice of reason giving and reason-evaluating, it will cease to be merely a tool and become a colleague" ([41], p. 48).

It would be useful at this point to distinguish between *instrumental rationality* and *reflective rationality* [42]. The former is about adopting suitable means to ends that are already fixed, while the latter is about choosing, evaluating, and reconsidering ends

[8] One area for future research into HMIs is with regards to the *division of labour* that occurs within joint actions; almost all joint actions are asymmetrical in the sense that the respective agents will be playing different roles in the activity. This has consequences not just for how these different roles have to be represented or understood by the respective agents, but also for the how moral responsibility is distributed across the partners (it is not always the case, in a joint action, that the partner who is causally responsible for the fatal mistake is the one who is morally responsible for it – think, for instance, of a joint action involving a parent and child).

themselves. It is the latter that we are claiming is not available to artificial systems. It is not available to artificial systems because it would require AGI, and, at least partly because of the frame problem, AGI is unlikely to come along in the near future, if at all [43].

Indeed, according to Stuart Russell, reflective rationality barely figures within the goals of 'current AI research': "[I]n both the logical-planning and rational-agent views of AI, the machine's objective – whether in the form of a goal, a utility function, or a reward function (as in reinforcement learning) – is specified exogenously" ([44], p. 328). This exogenous goal might be implicit in the design (i.e., inputted by the designer/manufacturer), or it might be added as input by the user. In either case, the end is not chosen by the machine.

If one is interacting with a machine, then, one is interacting with something that is potentially capable of highly sophisticated instrumental rationality but is incapable of reflective rationality. This, we propose, makes HMIs a different type of social interaction from human-human interactions. Explaining the role of the artificial system in a social HMI, we take the AI-Stance (where AI can stand for either artificial *intelligence* or artificial *intentionality*). The AI-stance is to be prepared to treat one participant of a social interaction – namely the artificial system – as completely arational in terms of reflective rationality (it is not able to compare or evaluate or reconsider ends), but to treat it as fully rational in terms of instrumental rationality. Indeed, we think it might go beyond this: unlike with human-human interactions, where expertise in ends and means tends to be matched (the expert in ends will tend to also be the expert in the means), in human-machine interactions there is an almost total mismatch: artificial systems might be highly expert in means, so much so that the human can completely defer to the machine in the means, but, when it comes to ends artificial systems are completely silent.

5.2. Implications for Moral Responsibility

As soon as certain artificial systems qualify as a new type of potential social interaction partners, the question arises to what extent they could then be attributed not only causal but also moral responsibility. Discussing conceivable ways of how moral responsibility could be distributed in social human-machine interactions, we investigate which features of artificial systems might speak for distributed responsibility (for more details see [45]).

For example, artificial agents can process and store a greater amount of data in a shorter time. This can have crucial consequences when examining the extent of their influence on the outcome of an interaction. The artificial system might be an expert in choosing certain means and human reaction times can simply be too slow to intervene effectively. Another question concerns the extent our limited ability to predict the behavior of artificial agents might absolve us from taking on a greater share of the responsibility. Consequently, one could argue that human interaction partners with reduced anticipation skills and a less developed ability to process and store data deserve a smaller share of responsibility in social human-machine interactions.

Evaluating artificial systems only through the lenses of a design stance, all we can attribute to them is causal responsibility. However, taking a full-blown intentional stance towards them the consequences regarding ascriptions of moral responsibility seem to contradict our common sense because it appears senseless to blame or forgive artificial systems and we have no idea how to punish them. Introducing an AI-stance, we can distribute moral responsibility in certain human-machine interactions within a limited domain. In other words, an AI-stance might be a solution to avoid too far-reaching

consequences of the project to describe (moral) agency as a gradual phenomenon, such as the close connection between responsibility and possible sanctions and the question of what status artificial systems have outside of social interactions. Nevertheless, future research should investigate the extent to which artificial systems could face up to their responsibilities if, for example, they were equipped with liability insurances.

6. Conclusion

Assuming that not all HMIs can appropriately be reduced to and described as mere tool-use, we claim that there are HMIs conceivable in which machines should be regarded as a new type of a social interaction partner.

Since this claim contradicts standard construals of agency and sociality we diagnose a terra incognita for which we have no established notions yet. A critical investigation of a Davidsonian-like approach and of a biological conception towards agency presents reasons motivating an approach utilizing gradual conceptions and minimal approaches. On the one hand we question a dichotomic distinction between mere behavior and agency because empirical results speak for a gradual development. On the other hand we suggest that relying on biological constraints might overlook the possibility that there are multiple realizations of agency possible.

Presupposing that one can describe social HMIs as a minimal asymmetric joint action, we explore the explanatory power of the epistemological stance terminology and argue that none of the three stances lead to an adequate explanation of social HMIs. This is why we suggest a further stance – the AI-stance – that is able to acknowledge that machines lack full-fledged agency but still have minimal agency. Thereby one can avoid reducing all HMIs to mere tool-use without being forced to ascribe full-fledged moral agency to such participants in social interactions.

References

[1] Duchomel P, Damiano L. Living with robots. Cambridge, MA: Harvard University Press; 2017. 262p.
[2] Stone P, Brooks R, Brynjolfsson E, Calo R, Etzioni O, Hager G, et al. Artificial intelligence and life in 2030. In: Proceedings of the One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Stanford, CA: Stanford University; 2016.
[3] Henschel A, Laban G. Cross ES. What makes a robot social? A review of social robots from science fiction to a home or hospital near you. Curr Robot Rep. 2021 (2): 9–19.
[4] Hornsby J. Simple-mindedness: In defense of naïve naturalism in the philosophy of mind. Cambridge, MA: Harvard University Press; 1997. 265p.
[5] Povinelli D, Vonk J. We don't need a microscope to explore the chimpanzee's mind. Mind Lang. 2006 19: 1–28
[6] Davidson D. Actions, reasons and causes. Journal of Philosophy. 1963 Nov;60(23): 685-99
[7] Davidson D. Agency. In: Binkley R, Bronaugh R, Marras, A., editors. Agent, action and reason. Toronto: University of Toronto Press; 1971. p. 3-37.
[8] Davidson D. Essays on actions and events. Oxford: Oxford University Press; 1980. 304p.
[9] Davidson D. Rational animals. Dialectica. 1982 36: 317-28.
[10] Davidson D. Inquiries into truth and interpretation. Oxford: Oxford University Press; 1984. 296p.
[11] Davidson D. Subjective, intersubjective, objective. Oxford: Oxford University Press; 2001. 237p.
[12] Lepore E. Ludwig K. Donald Davidson: Meaning, truth, language, and reality. Oxford: Oxford University Press; 2005. 466p.
[13] Lepore E. Ludwig K. Donald Davidson's truth-theoretic semantics. Oxford: Oxford University Press; 2007. 362p.

[14] Perler D, Wild M. Der Geist der Tiere – Philosophische Texte zu einer aktuellen Diskussion. Frankfurt: Suhrkamp; 2005.
[15] Premack D, Woodruff G. Does the chimpanzee have a theory of mind? Behavioral Brain Sciences 1978; 1, 515-526.
[16] Heyes C. False belief in infancy: a fresh look. Developmental Science 2014; 17 (5), 647-659.
[17] Heyes C. Animal mindreading: what's the problem? Psychonomic Bulletin & Review 2015; 22 (2), 313-327.
[18] Vesper C, Butterfill S, Knoblich G, Sebanz, N. A Minimal Architecture for Joint Action. Neural Networks 2010; 23, 998-1003.
[19] Warneken F, Chen F, Tomasello M. Cooperative activities in young children and chimpanzees. Child Dev. 2006; 77, 640-663.
[20] Perner J. Understanding the representational mind. Cambridge, MA: MIT Press; 1991. 348p.
[21] Tomasello M. Origins of human communication. Cambridge MA: MIT Press; 2008. 408p.
[22] Sterelny K. The evolved apprentice: How evolution made humans unique. Cambridge, MA; MIT Press; 2014. 242p.
[23] Henrich J. The secret of our success: How culture is driving human evolution, domesticating our species and making us smarter. Princeton: Princeton University Press; 2016. 445p.
[24] Searle J. Minds, Brains and Science: The 1984 Reith lectures. London: British Broadcasting Corporation; 1984. 99p.
[25] Strasser A. Kognition künstlicher Systeme. Berlin: De Gruyter; 2006. doi: 10.1515/9783110321104.
[26] Strasser A. Can artificial systems be part of a collective action? In: Misselhorn C, editors. Collective Agency and Cooperation in Natural and Artificial Systems. Explanation, Implementation and Simulation. Berlin: Springer; 2015. Philosophical Studies Series, 122, 205-218. doi: 10.1007/978-3-319-15515-9_11.
[27] Strasser A. From tools to social agents. Rivista Italiana di Filosofia del Linguaggio 2020; 14 (2), 76-87, doi: 10.4396/AISB201907.
[28] Wellman H, Cross D, Watson J. Meta-Analysis of Theory of Mind Development: The Truth About False-Belief. Child Development 2001; 72 (3): 655-84.
[29] Fiebich A., editor. Minimal cooperation and shared agency. Switzerland: Springer Books; 2020. 217p.
[30] Pacherie E. Intentional joint agency: Shared intention lite. Synthese 2013; 190 (10), 1817-1839.
[31] Butterfill S, Apperly I. How to construct a minimal theory of mind. Mind and Language 2013; 28 (5), 606-637.
[32] Michael J, Sebanz N, Knoblich G. The Sense of Commitment: A Minimal Approach. Frontiers in Psychology 2016; 6, 1968.
[33] Dennett D. Intentional systems. Journal of Philosophy; 1971 Feb 68(25): 87-106.
[34] Dennett D. The intentional stance. Cambridge, MA: The MIT Press; 1987.
[35] Behdadi D, Munthe C. A normative approach to artificial moral agency. Minds and Machines 2020; 30: 195-218.
[36] Neuhäuser C. Some sceptical remarks regarding robot responsibility and a way forward. In: Misselhorn C, editor. Collective action and cooperation in natural and artificial systems. Switzerland: Springer; 2015. p. 131–46.
[37] Zadwinksi T. The many roles of the intentional stance. In: Heubner B, editor. The philosophy of Daniel Dennett. Oxford: Oxford University Press; 2018. p. 36-56
[38] Blomberg O. Shared intention and the doxastic single end condition. Philosophical Studies; 2016.173 (2):351-372.
[39] Searle J. Collective intentions and actions. In: Cohen PR, Morgan J, Pollack M, editors. Intentions in communication. Cambridge, MA: MIT Press; 1990. p. 401-15.
[40] Bratman M. Shared cooperative activity. Philosophical Review; 1992 101(2): 327-41.
[41] Dennett D. The age of post-intelligent design. In: Gouvia SS, editor. The age of post-intelligent design. Delaware: Vernon Press; 2020. p. 27-62.
[42] Schmidtz D. Choosing ends. Ethics; 1994 Jan 104(2): 226-51.
[43] Marcus G, Davis E. Rebooting AI: Building artificial intelligence we can trust. New York: Pantheon Books; 2019. 290p.
[44] Russell S. Artificial intelligence: A binary approach. In: Liao SM, editor. Ethics of artificial intelligence. Oxford: Oxford University Press; 2020. p. 327-41.
[45] Strasser A. Distributed responsibility in human-machine interactions. AI and Ethics 2021. doi: 10.1007/s43681-021-00109-5 2021.