Author's copy. Appeared in: Nørskov, M., Seibt J., Quick O. 2020. Culturally Sustainable Social Robotics—Proceedings of Robophilosophy 2020. Series Frontiers of AI and Its Applications, IOS Press, Amsterdam.

Towards a Middle-Ground Theory of Agency for Artificial Intelligence

Louis LONGIN[a,1]
[a]*Department of Philosophy, Ludwig-Maximilians-University, Germany*

Abstract. The recent rise of artificial intelligence (AI) systems has led to intense discussions on their ability to achieve higher-level mental states or the ethics of their implementation. One question, which so far has been neglected in the literature, is the question of whether AI systems are capable of action. While the philosophical tradition appeals to intentional mental states, others have argued for a widely inclusive theory of agency. In this paper, I will argue for a gradual concept of agency because both traditional concepts of agency fail to differentiate the agential capacities of AI systems.

Keywords. Artificial agency, gradual agency, artificial intelligence, robotics

1. Introduction

With the emergence of complex artificial intelligence (AI) and advanced robotics, AI has been applied in various industries from automotive to healthcare and robotics. AI is now widely used to aid human decision making in medical diagnostics, to facilitate autonomous driving and to analyse complex data structures. Modern robotic AI-systems such as Joyforall's robotic Companion Pet or Hanson Robotics' SOPHIA [1] are even often treated by humans as interactive agents and are expected to behave as such. This broad understanding of agency is supported by most of cognitive and computational science (see [2]). However, the philosophical tradition based on Anscombe, Bratman and Davidson denies any attribution of agency to such computational systems. With these different theories of agency, the question emerges: under which conditions, if at all, can robots and AI systems count as agents?

Agency in the narrow sense has been promoted mainly by the philosophical tradition [3]. It links agency, defined as the capacity to perform intentional actions, strongly to human-like mental states. Davidson, for instance, argues that something counts as an action only if it is done intentionally, for reasons and is caused by the right mental states of beliefs and desires in the right way. The advantage of requiring mental states for action is that it provides good grounds for the intuitive difference between behaviour and action. Hence, the event-causal theory of action has been accepted as the standard theory of agency up until now [2].

Agency in the broad sense captures a wide range of intuitions on attributing agency to non-human systems. According to dynamic system theory [4], systems are defined as

[1] Corresponding Author: Louis Longin, Ludwig-Maximilians-University, Faculty of Philosophy, Philosophy of Science and the Study of Religion, Geschwister-Scholl-Platz 1, 80359 Munich, Germany; e-mail: louis.longin@campus.lmu.de.

agents merely in virtue of their behaviour interaction with the environment. Beer's definition is picked up by Barandiaran et al. who unify these various intuitions and propose a minimal theory of agency [5]. Here, a system is considered as an agent if the system is defined by itself (individuality condition), is capable of actively regulating its environmental interactions (interactional asymmetry condition) and does so according to some internal norms and goals (normativity condition). On this conception, even a biological cell can be considered as an agent.

The challenge for applying a theory of agency to robots and AI systems resides mainly in their wide-ranging interactive capacity: from data analysis software to Terminator-like robots. While robotic vacuum cleaners, for instance, can tidy your home, they lack any linguistic expression found in natural language models, task flexibility found in meta-learning systems or human-like behaviour found in humanoid robots.

In this paper, I will show that neither the broad nor the narrow notion of agency can successfully capture the different agential capacities found in robotic and AI systems. Instead, I propose a gradual account of agency that postulates different kinds of agency based on varying criteria of agential capacities. A graduate account of artificial agency has strong implications to connected debates like ethical and legal responsibility. If an artificial system counted as an agent, then they could be moral agents and bear responsibility for their actions.

2. The narrow notion of agency

2.1. Davidson's event-causal framework

The common philosophical framework under which agency is understood represents Davidson's event-causal theory of action. It holds that something counts as an action if it is done intentionally, for reasons and is caused by the right mental states of beliefs and desires in the right way [6]. This represents a reductive conception of agency according to which occurrences of agency can be reduced to pairs of agent-involving mental states and events.

Davidson's theory of action and agency has been developed over decades in various essays. In particular, his essays on *Actions, Reasons, and Causes* (1963), *Agency* (1971) and *Intending* (1978) ground his event-causal framework. An agent's role is reduced to the causal roles of agent-involving states and events [7]. In his essays, he analyses the concept of action and develops a causal explanatory framework for action. Any explanation of an action is given by the agent's reasons for acting [8]. Generally, Davidson provides a reason-based explanation of action in terms of mental states which cause action and make the action intelligible to the agent and others through the process of rationalisation.

Davidson's framework holds that, for every action, something can be said to justify the action from the agent's perspective. Mental states rationalise an action if they can causally explain the performance of the action. In particular, those mental states which rationalise an action are the right mental states [9]. In particular, Davidson highlights pro attitudes, intentions and beliefs as these relevant mental states. Pro attitudes describe mental states like desires, inclinations or urges, which are formed towards a particular kind of action whereas beliefs represent the individual outlook and understanding of the world. Davidson holds that acting with an intention precedes having a pro attitude towards the respective action. He changes his position on the reducibility of intentions

in the explanatory role of action from reducible to desires and beliefs in his early work to irreducible in later work. Intentions thus play an essential role in practical reasoning and guidance of action.

The advantage of the traditional, narrow conception of agency is that it clearly distinguishes between an intentional action, as behaviour that is produced by certain mental states, and behaviour that is explained only in reference to a purely material, causal relation [10]. In other words, a narrow account of agency can distinguish human action, based on mental states, from non-human behaviour that applies to animals and artefacts.

2.2. Application to AI systems

Before applying a conception of agency to robotic and AI systems, it must be clear what the concept is applied to. Generally, there are three different kinds of artificial systems: those that are designed by human engineers from the top-down, those that adapt to their environment through self-learning and their hybrid combination [11, 12].

Top-down systems are typically rigid systems that follow specified behaviour rules. Those rules are determined by the human developer and guide the system's interaction and decision making. Hence, top-down systems operate in a very specific application domain on a high competence level. Technically, a top-down system is also often identified as a symbolic system because it operates on well-defined symbols and logical reasoning. Expert systems are a typical example for top-down systems, where expert-level, human knowledge is implemented to facilitate process automation, planning or predictions. This includes ethical decision recommendation systems such as the logical formalisation of Kantian ethics [13, 14] or process planning systems such as Comex [15] and Cakes-Ists [16].

Bottom-up systems, in contrast, learn to make decisions by themselves [17,18]. What is given by the human developers is the learning algorithm as well as the application environment instead of the explicit specification of a set of behaviour rules. A bottom-up system utilises machine learning to develop its internal representations of the world around it to solve a given task. Technically, instead of symbolic computation, bottom-up systems use sub-symbolic computation such as artificial neural networks and evolutionary systems. Popular examples include case-based reasoning systems like Truth-Teller or Sirocco [19] and moral analytic machine learning systems [20].

Lastly, hybrid architectures are commonly inspired by models of human cognition and seek to implement forms of human-level cognitive functions by combining the self-learning capacities of bottom-up systems with explicit cognitive structures of top-down systems [12]. This allows hybrid architectures such as LIDA [18] or ACT-R [21] at least in principle to emulate human-like behaviour (for review see [22]).

When considering whether robots and AI systems can fulfil the criteria of the narrow notion of agency, the discussion turns towards whether artificial systems have or could have mental states like beliefs, desires and intentions. Following Davidson's event-causal theory of action, in order for an AI system to be considered as a human-like agent, it must initiate something which is caused by intentions, the right pro-attitudes and desires which represent the reasons for the possible action.

In a more general sense, in order to count as an agent, the artificial system must have some intentionality. In other words, it must possess the ability to represent, i.e. be about things and properties of the internal or external world. Only with intentionality, an entity can have the, for agency required, mental states.

The kind of intentionality in question is what [13] describes as internal intentionality, i.e. intentionality that emerges from the system itself. Internal intentionality stands in contrast to external intentionality, which is intentionality attributed to the system from external, behavioural observations or externally scripted output. Some examples for external states are speech acts, maps or basic interactive robots. Here, the produced symbols are intentional and represent something. However, these representations are not emerging from the system itself but are rather implemented by human developers. Just as Searle's Chinese Room Argument shows for semantics [24], computational systems require external input to produce meaningful, intentional representations.

While ascriptions of external intentionality are key for attributive accounts of mental states and agency, such as the intentional stance [25, 26], internal intentionality is necessary for having mental states and being an agent in the narrow sense. Only with inherently emerging intentional states artificial systems can satisfy the criterion of agency laid out by Davidson. Under the Davidsonean framework, something counts as an action only if it is done intentionally, for reasons and is caused by beliefs and desires in the right way. Reasons and 'mental' states that are given externally do not cause an action in the same way internal reasons and mental states do. Such properties are necessarily emergent as they are located neither in hardware nor in software [13]. These emergent properties exceed the capabilities of current top-down, bottom-up and hybrid systems, and it remains unclear how such intentional states could develop.

3. The broad notion of agency

3.1. Minimal agency

The broad conception of agency does not restrict agency to intentional action but instead uses a wide scope to capture the common intuition of attributing agency to other objects. In this broad sense, agency is everywhere and is roughly understood as the manifestation of a capacity to initiate interaction with the environment in pursuit of some goal. Understanding agency in a broad sense encompasses many different intuitions about ascribing a pre-critical concept of agency to non-human systems.

In the very basic sense, a broad notion of agency is based on an observed causality between a behaving system and an occurring event in the environment. Recent artefacts like Microsoft's Twitter chatbot Tay, which became racist after less than 24 hours from its release [27], or VW's emission regulating software, which manipulated car emissions, show that artefacts nowadays can leave an independent causal impact on the world around them. If agency is understood merely in terms of a causal interaction with the environment, then agency dramatically extends the scope of possible agents beyond the human agents to include any kind of interactive system.

Formalising the approach of agency in a broad sense means to consolidate the various intuitions within a general theory of agency that applies to natural as well as artificial agents. Such an approach has been provided by [5] with the theory of minimal agency. In particular, [5] identify three necessary and sufficient conditions for a working concept of agency: *individuality*, *interactional asymmetry* and *normativity*.

Individuality, as the first criterion of minimal agency, points out that in order to distinguish between an agent and the environment, an agent must be individually identifiable. This allows for the development of any kind of relationship between the agent and objects in the environment. Any agent possesses some form of identity which

allows to separate itself from and dynamically interact with the environment. This distinction can be conscious as well as unconscious but represents a prerequisite for any interaction.

Crucial here is that the individuality condition is not based on the judgment of an external observer. Any such dependency would require another justification of the individuality of the observer leading to an infinite regress. Hence the agent-environment distinction and determination of individuality must be internal. Only this capacity allows a system to form in relations with other objects in the environment.

Interactional asymmetry, as the second criterion of minimal agency, formalises the necessary interaction between an agent and its surroundings. An agent is always the source of an action and modulates the environment by itself to suit its needs. The relationship between an agent and the environment is hence asymmetric as the agent is capable of asserting itself on the environment by modulating some parametrical conditions of the structured relation between itself and the environment. It is further possible but not necessary for an agent to act upon this capability. Then an agent not only is capable of producing a change of some environmental parameters but also actively modulates them to achieve some particular outcome.

Normativity, as the third criterion of minimal agency, is necessary to rule out any random interaction with the environment and ensure that the action in question is in line with the endorsed goals and norms of the system. Any interactive modulation of environmental condition represents a modulation to satisfy a given norm or goal. Such norms are not given by the environment but are rather generated by the system itself. Agents fundamentally regulate their interaction with the environment based on their perceived success or failure of fulfilling their internalised norms, which allows the system to distinguish between different outcomes of its actions.

3.2. Application to AI systems

While [5] are themselves sceptical about the application of minimal agency to artificial systems, I believe that their criteria can indeed be satisfied by AI systems.

According to the first *criterion of individuality*, an agent must be sufficiently distinguishable from its environment. From the perspective of an external observer, all three kinds of AI systems can be distinguished from their environment through their hardware or software implementation. From the system's perspective, this condition also holds because each system is programmed to interact with the environment which necessitates the system's ability to distinguish itself from the environment.

According to the second *criterion of interactional asymmetry*, a minimal agent must be the source of an active modulation of its coupling with the environment. This can take place on two main interpretations. On the energetic interpretation, a minimal agent must expend energy in order to modulate the environment in some form. All three kinds of AI systems satisfy the criterion of interactional asymmetry because they expand computational resources in order to interact with the environment. On the statistical interpretation, a minimal agent must be the statistically significant reason for a change in the environment. While top-down systems have troubles with this interpretation, under a normal scope, all three kinds of AI systems also satisfy the criterion of interactional asymmetry according to the statistical interpretation.

According to the third *criterion of normativity*, a minimal agent must pursue some underlying goal which guides its environmental interactions. This trivially holds for all three kinds of AI systems which are all programmed to fulfil a certain goal in the form

of error-minimisation, reward-maximisation or successful execution of its given internal function.

This means that each AI system satisfies the three conditions of minimal agency and can be considered a minimal agent under the broad conception of agency.

4. The gradual notion of artificial agency

4.1. The problem

The reason to explore the grounds for a middle ground theory of agency is that neither established theory of agency can distinguish between the intuitively perceived difference in agential capacities of artificial systems. On the one side, the narrow notion of agency under-generates the class of possible agents by imposing overly strict criteria for action. Here, no artificial system counts as an agent because it lacks any internally emergent intentional states. On the other side, the broad notion of agency over-generates the class of possible agents by imposing overly loose criteria for action. Here, all kinds of artificial systems count as agents as they fulfil the minimal criteria of agency.

These two extreme notions of agency do not capture the intuitively perceived differences between, for instance, humanoid robots like Sophia diagnostic software systems. These intuitions are mirrored in other fields of non-human agency like animal [28–31] and collective agency [32–34]. In both cases, the narrow notion of agency cannot capture the agential capacities of these non-human agents, which has led researchers to develop accompanying kinds of agency valid for their domain. Furthermore, we might have reasons to argue that even humans do not always act according to the narrow, event-causal model of agency because most of our daily actions are, in fact, automatic and subconscious. Furthermore, in other cases of illusions or for children, we might still be inclined to attribute cases of agency while, on the traditional conception of agency, the necessary mental states are missing or not related to the action in the right way [35, 36].

Similarly, a theory of agency for robotic and AI systems must accommodate the intuitively perceived difference in agential capacity for those systems. An overly loose theory of agency - similarly to an overly strict theory of agency - cannot capture these differences.

4.2. The solution

What can fill this conceptual void is to understand agency of artificial systems as a gradual concept. A gradual notion of agency maps various degrees of agential capacities to different kinds of agency. The more capable a system is, such as through the possession of intentional states, the more demanding criteria of agency the system can fulfil. The gradual notion of agency can be thought of as a discrete scale with the broad notion of agency on the one side, the narrow notion of agency on the other and additional sets of criteria in the middle.

The gradual notion of agency can capture the intuitively perceived difference in agential capacities in artificial systems by making room for kinds of agency that lie between the traditional narrow and broad notions of agency. One candidate for providing such middle-ground criteria of agency stands out: the minimal theory of mind. Because the minimal theory of mind explains a wide range of animal and child behaviour without

presupposing mental states [38], it can bridge the gap between a demanding but not overly restrictive account of agency.

Another possible kind of agency that could be similarly successful in differentiating the agential capacities of artificial systems is animal agency. Animal agency requires a physical body, some form of subjectivity, attributable elementary intentional states as well as the initiation of movement [30, 31]. However, due to the length of this paper, I will focus only on the minimal theory of mind. My intuition is nonetheless that an account of animal agency can provide an additional kind of middle-ground agency that can also be applied to advanced artificial systems.

The minimal theory of mind (mToM), as proposed by [38], aims to provide a theory of mind which is rich enough to explain the systematic success of solving cognitive tasks while also not requiring propositional attitudes or other kinds of representations. By focusing on low-level cognitive capacities as well as limited cognitive resources, this theory has been successfully used to explain the success in belief tracking tasks for children and primates [38, 39]. The minimal theory of mind rests on five principles.

The first principle holds that a mToM agent must be able to track the function of objects by linking their observed movements to their inherent goals. The agent does not need to understand the action but pick out at which goal the action might be directed. The second principle states that a system cannot act on an object without having encountered it. This includes having a field of awareness in which objects can be encountered depending on the object's proximity, orientation, lighting etc. towards the agent. Thirdly, in order to perform a successful action, the mToM agent has to register the object by linking the encountered object to its conditions of location, orientation etc. Correct registration represents the condition for successful action. The fourth principle holds that, given a previous object registration, the mToM agent will act as if the object were in the location previously registered. For example, in a false belief task, the false belief can only be predicted if the belief is registered under the occurring conditions and is then contrasted with the changing environment [40, 41].

The minimal theory of mind allows explaining children's success in false belief tasks who do not possess fully developed cognitive capacities. Inspired by [42], it seems possible to ground some form of minimal, contentless intentionality based on the mToM. As children use teleological behaviour reading instead of reflective, intentional mental states to track other's beliefs, they do not possess full, adult-like intentionality but rather rely on a minimal form of intentionality in terms of directness and responsiveness to track other's beliefs.

Applying the minimal theory of mind and the respective idea of minimal intentionality to artificial systems, it becomes clear that not many systems can satisfy the criteria for a mToM-inspired middle-ground notion agency. Recall that for mToM the respective agent must be able to track the teleological function of observed objects, to encounter objects within a field of awareness, to register the agent-object relation at the time of the encounter and to track that relation over time.

Common top-down systems do not classify as mToM agents because their underlying code is designed to fulfil an externally determined task like issuing recommendations or solving domain-specific problems. Similarly, bottom-up systems might learn to solve a given task, but they do so by extrapolating data patterns through iterative training sessions. In order for any top-down or bottom-up system to be a mToM agent, they must be specifically designed to attribute teleological functions by registering their system-object relation as part of their environmental interaction.

One approach that has implicitly attempted to go in that direction but ultimately falls short is the belief-desire-intention (BDI) model of artificial systems. Here, a model of human practical reasoning, originally developed by [43], is combined with artificial reasoning systems [44] to model interaction within a multi-agent system. While there are many different implementations, a BDI system generally tracks states of the world by recording its relation to the environment and possesses an inherent drive to achieve its given goals. While these systems can cooperatively interact with other agents, they do so in virtue of basic environmental observation without any goal-state attribution. What is missing in these top-down BDI agents is a bottom-up machine learning component that attributes teleological functions to observed objects. Such prospective hybrid systems can then arguably be considered mToM agents.

5. Discussion

When applying each theory of agency to AI systems, it also becomes clear that AI agency can hardly be identified with either theory alone. Agency in the narrow sense is too restrictive because it tries to apply the concept of human agency in terms of mental states and intentional action to artefacts. Agency in the broad sense, on the other hand, is too general because it attributes the same level of agency to systems which are fundamentally distinct in their internal functioning and agential capabilities. Thus, it seems like agency for AI systems requires a conception of agency which is more general than agency in the narrow sense but more restrictive than agency in the broad sense. Such a middle ground does not eliminate the requirement of intentionality altogether but instead focuses on the possession of particular internal mental states.

An adequate conception of agency for AI systems should be sensitive to the different agential capacities in AI systems while also not promoting overly strict criteria of agency. Therefore, it is necessary to reframe the debate of agency from the traditional categorical understanding of agency to a *gradual conception of agency*. A gradual conception of agency could integrate the broad and the narrow notions of agency into one framework of agency as well as provide the conceptual room for a middle ground conception of AI agency. An entity can then be classified as an agent according to differently strict criteria of agency. The narrow notion of agency represents the highest attainable level of agency by providing the strictest set of criteria for action. The broad notion of agency, on the other hand, defines the lowest possible level of agency by advocating the basic set of criteria for action. The middle ground between the broad and the narrow notion of agency can consist of different levels of partial agency which each employ stricter criteria of agency when compared to the broad notion but looser criteria when compared to the narrow notion.

This gradual conception of agency has various advantages over the traditional categorial conception. On the one hand, it allows for the integration of both opposing theories into one conceptual framework. This alleviates the pressures for both theories by adapting their criteria of agency to capture a wider or more restricted range of potential agents. On the other hand, it can provide conceptually rich criteria for an account of artificial agency by analysing the agential capacities of such systems. This way, it would be possible to classify AI as low or high-level agents based on the development of their internal capacities. A promising approach towards such a comprehensive notion of AI agency could focus on the internal functioning and capabilities of the AI systems. The internal functioning of a system describes how AI

systems compute, interact and learn through their interactions with the environment. Each instance of top-down, bottom-up and hybrid AI systems can possess different sensors, reasoning structures and learning algorithms which fundamentally inform what those systems are capable of. Top-down systems, for example, are bound by their implemented rules and top-down principles to interact with the environment in some determinate way. Bottom-up systems, on the other hand, are only given a reward structure and a self-learning algorithm and have to develop their kind of interaction on their own. Both instances operate functionally very differently, which should be reflected in their conception of agency.

References

[1] Goertzel B, Mossbridge J, Monroe E, Hanson D, Yu G. Humanoid robots as agents of human consciousness expansion. [Preprint] 2017. Available from: https://arxiv.org/abs/1709.07791.
[2] Schlosser M. Agency. In: Zalta EN, editor. Stanf Ency Philos. Stanf Uni: Metaphysics Research Lab; 2019.
[3] Davidson D. Actions, reasons, and causes. J Philos. 1963;60(23):685–700.
[4] Beer RD. A dynamical systems perspective on agent-environment interaction. Artif Intell. 1995;72(1-2):173–215.
[5] Barandiaran XE, Di Paolo E, Rohde M. Defining Agency: Individuality, Normativity, Asymmetry, and Spatio-temporality in Action. Adapt Behav. 2009;17(5):367–86.
[6] Davidson D. Essays on actions and events. Oxford: Clarendon Press; 2001. Chapter 3: Agency. p.43-62
[7] Schlosser ME. Agency, ownership, and the standard theory. In: Auilar JH, Buckareff AA, Frankish K, editors. Waves Philos Action. London: Palgrave Macmillan; 2011, p. 13–31.
[8] Davidson D. The essential Davidson. New York: Oxford University Press; 2006.
[9] Davidson D. The essential Davidson. New York: Oxford University Press; 2006. Chapter 1: Actions, reasons, and causes. p.23-36.
[10] Johnson DG, Verdicchio M. AI, Agency and Responsibility: The VW Fraud Case and Beyond. AI Soc. 2018;34(3):639–47.
[11] Wallach W, Allen C, Smit I. Machine morality: Bottom-up and top-down approaches for modelling human moral faculties. AI & SOCIETY. 2007;22(4):565–82.
[12] Allen C, Smit I, Wallach W. Artificial morality: Top-down, bottom-up, and hybrid approaches. Ethics Inf Technol. 2005;7(3):149–55.
[13] Powers TM. On the moral agency of computers. Topoi. 2013;32(2):227–36.
[14] Powers TM. Prospects for a Kantian machine. IEEE Intell Syst. 2006;21(4):46–51.
[15] Grahovac D, Devedzic V. COMEX: A cost management expert system. Expert Syst Appl. 2010;37(12):7684–95.
[16] Lee KC, Lee S. A causal knowledge-based expert system for planning an Internet-based stock trading system. Expert Syst Appl. 2012;39(10):8626–35.
[17] Allen C, Varner G, Zinser J. Prolegomena to any future artificial moral agent. J Exp and Theor Artif Intell. 2000;12(3):251–61.
[18] Wallach W, Franklin S, Allen C. A conceptual and computational model of moral decision making in human and artificial agents. Top Cogn Sci. 2010;2(3):454–85.
[19] McLaren B. Lessons in machine ethics from the perspective of two computational models of ethical reasoning. Proceedings of the 2005 AAAI Fall Symposium on Machine Ethics; Menlo Park, California: AAAI Press; 2005.
[20] Conitzer V, Sinnott-Armstrong W, Borg JS, Deng Y, Kramer M. Moral decision making frameworks for artificial intelligence. Proceedings of the Thirty-first AAAI conference on Artificial Intelligence; San Francisco: AAAI Press; 2017.
[21] Anderson JR. ACT: A simple theory of complex cognition. Am Psychol. 1996;51(4):355.
[22] Kotseruba I, Tsotsos JK. 40 years of cognitive architectures: Core cognitive abilities and practical applications. Artif Intell Rev. 2020;53(1):17–94.
[23] Crane T. Intentionality as the Mark of the Mental. Royal Institute of Philosophy Supplement, Cambridge University Press. 1998;43:229–51.
[24] Searle JR. Minds, brains, and programs. Behav Brain Sci. 1980;3(3):417–24.
[25] Dennett DC. The Intentional Stance. Cambridge, London: MIT Press; 1989.
[26] Dennett DC. Intentional systems. J Philos. 1971;68(4):87–106.

[27] Beran O. An Attitude Towards an Artificial Soul? Responses to the "Nazi Chatbot". Philos Investig. 2018;41(1):42–69.
[28] Delon N. Animal agency, captivity, and meaning. Harv Rev Philos. 2018;25:127-46.
[29] Jamieson D. Animal agency. Harv Rev Philos. 2018;25:111–26.
[30] Steward H. A Metaphysics for Freedom. Oxford: Oxford University Press; 2012.
[31] Steward H. Animal agency. Inquiry. 2009;52(3):217–31.
[32] List C. Group Agency and Artificial Intelligence [Preprint] 2019. Available from: http://philsci-archive.pitt.edu/15980/.
[33] Stapleton M, Froese T. Is collective agency a coherent idea? Considerations from the enactive theory of agency. In: Misselhorn C, editor. Collective agency and cooperation in natural and artificial systems; Cham: Springer; 2015. p. 219–36.
[34] Tenenbaum S. Representing Collective Agency. Philos Stud. 2015;172(12):3379–86.
[35] Clarke R. Libertarian accounts of free will. New York: Oxford University Press; 2003.
[36] James A. Agency. In: Qvortrup, J, Corsaro, W, Honig, M, editors. The Palgrave Handbook of Childhood Studies, Springer; 2009, pp. 34–45.
[37] Sebo J. Agency and moral status. J Moral Philos. 2017;14(1):1–22.
[38] Butterfill SA, Apperly IA. How to construct a minimal theory of mind. Mind Lang. 2013;28(5):606–37.
[39] Wimmer H, Perner J. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition. 1983;13(1):103–28.
[40] Onishi KH. Do 15-Month-Old Infants Understand False Beliefs? Science. 2005;308(5719):255–8.
[41] Hare B, Call J, Tomasello M. Do chimpanzees know what conspecifics know? Anim Behav. 2001;61(1):139–51.
[42] Hutto DD, Satne G. The Natural Origins of Content. Philosophia. 2015;43(3):521–36.
[43] Bratman ME, Intention, Plans, and Practical Reason. Cambridge: Harvard University Press; 1987.
[44] Georgeff M, Pell B, Pollack M, Tambe M, Wooldridge M. The Belief-Desire-Intention Model of Agency. In: Müller JP, Rao AS, Singh MP, editors. Intelligent Agents V: Agents Theories, Architectures, and Languages, Berlin, Heidelberg: Springer; 1999, p. 1–10.
[45] Asaro PM. What should we want from a robot ethic. Int Rev Inf Ethics. 2006;6:9–16.