Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Behavioral Sciences

Synthetic agency: sense of agency in artificial intelligence

Roberto Legaspi[1,2], Zhengqi He[1,2] and Taro Toyoizumi[1,2]

The concept of sense of agency (SoA) has garnered considerable attention in human science at least in the past two decades. Coincidentally, about two decades ago, artificial intelligence (AI) research witnessed an explosion of proposed theories on agency mostly based on dynamical approaches. However, despite this early burst of enthusiasm, SoA models in AI remain limited. We review the state of AI research on SoA, seen predominantly in developmental robotics, *vis-à-vis* the psychology and neurocognitive treatments, and examine how AI can further achieve stronger SoA models. We posit that AI is now poised to better inform SoA given its advances on self-attribution of action–outcome effects, action selection, and Bayesian inferencing, and argue that synthetic agency has never been more compelling.

Addresses
[1] Laboratory for Neural Computation and Adaptation, RIKEN Center for Brain Science, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan
[2] RIKEN CBS-OMRON Collaboration Center, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan

Corresponding author: Toyoizumi, Taro (taro.toyoizumi@riken.jp)

Current Opinion in Behavioral Sciences 2019, **29**:84–90

This review comes from a themed issue on **Artificial intelligence**

Edited by **Matt Botvinick** and **Sam Gershman**

https://doi.org/10.1016/j.cobeha.2019.04.004

Introduction

Sense of agency (SoA) has become increasingly significant in philosophy, psychology, legal ethics and the cognitive neurosciences [1,2••]. SoA is the subjective experience that oneself initiates and controls its own actions and, through them, the external world [2••]. Hence, SoA grounds our sense of self, distinct from the world, and all kinds of causally efficacious self-world interactions mediated by our intentional actions. SoA has been posited fundamental to the experience of volition, self-awareness because of its self-other distinction, understanding the causal structure of the world, and the social concept of responsibility for one's own actions (see Ref. [3•]). It is therefore not surprising that SoA has also attracted the attention of researchers in artificial intelligence (AI) who aim to build autonomous, self-aware artifacts capable of purposeful actions [4,5].

Here lies the quandary. We strongly desire that we control the technologies we use [6]. However, SoA in AI would permit AI to have a subjective recognition of its own agency. It may be the case that an AI with high SoA is perceived as deterrent to our own SoA because we feel we are being controlled by the AI. Hence, this dilemma of joint human and synthetic agency might as well be front and center of our discourse on human–AI interaction: the pivotal point being an AI with high level of control leads to a dystopic future.

It was the rise of behavior-based robotic AI in the 1990s that catalyzed an outbreak of proposed theories on agency that were mostly based on dynamical approaches [4], but distant from the notions of subjective experience, cognition and intentionality. It is only recent that SoA, with human science theoretical underpinnings, is being adopted in AI. Consequently, concrete implementations are limited so far. Most can be seen in cognitive developmental robotics [7], where the robot distinguishes itself from the world to enhance its motor and cognitive skills through sensorimotor predictive processes [8•,9]. However, the human science investigation of SoA reveals that a full account of SoA should also consider non-sensorimotor cues (e.g. background beliefs and environmental cues) and ad hoc reasoning [10,11••].

We argue that AI research should rethink its treatment of SoA and go beyond its current reach that only provides incomplete forms of SoA. We put forward a simple, but coherent, synthesis of key theoretical human science treatments of SoA, and use this synthesis to make evident the current state of AI research on SoA and why it is lacking. However, we also believe that AI research already has in its disposal the techniques and tools to implement a more robust SoA. Hence, as we map pertinent AI components to major constituents of the synthesis, it becomes evident that AI research is actually poised to inform stronger models of SoA.

Synthetic agency in human–AI interaction

We mentioned that human and synthetic agencies to coexist harmoniously is a non-trivial problem. In instances where the AI overrides human control, what then becomes of human SoA? This agentive ambiguity in human–AI interaction, which could degrade human SoA, presents interesting challenges. We suggest the answer lies in a two-pronged perspective of SoA in both human and AI (Figure 1): *first*-person and *second*-person perspectives (*Fp* and *Sp*, respectively). In *Fp*, the AI (or

human) possesses SoA intrinsically, that is, it could infer if the outcome is caused by itself or someone (something) else. In *Sp*, the AI (or human) could perceive the other's SoA through its model or mental representation of the other's *Fp*. Humans can be very liberal in ascribing to artifacts that have no genuine intrinsic intelligence certain mental states or behavioral features as long as they demonstrate minimal human features or adhere to a rule of human social interaction (so-called intentional stance [12] and anthropomorphism [13]). On the other side, AI approaches for recognizing human intentions [14,15] and actions [16], as well as inferring action–outcome causal relations [17] exist. However, further advance would be to quantitatively assess and manipulate human SoA. For example, the AI could change its level of support to the human adaptively, that is, to achieve its intention without imposing, if human SoA decreases. For example, it has been shown in Ref. [18] (see also Ref. [19]), albeit not AI research, how human pilot's SoA can be manipulated by a flight simulator by varying its levels of automation, and measures both explicitly and implicitly the pilot's SoA relative to changes in automation level.

A goal is for AI to understand and adapt its *Fp* in order to enhance, rather than undermine, human-*Fp* (Figure 2). The dilemma is that by demonstrating a high *Fp* the AI influences human-*Sp*, which consequently causes human-*Fp* to decrease (Figure 2a). To resolve this, when AI-*Sp* is influenced by the low human-*Fp*, the AI must then carefully choose its next actions, or make no action at all, to heighten human-*Fp* (Figure 2b). The AI, with an understanding of its own *Fp*, can then better predict the dynamics of human-*Fp* and efficiently increase it (Figure 2c). Our main concern in this opinion piece is AI-*Fp*. In the same way, for example, that although AI can recognize and empathetically respond to human emotions [20], some researchers continue to raise the prospect of AI having actual emotions to better interact with humans [21]. We posit that with *Fp* an AI can better predict and respond to human SoA.

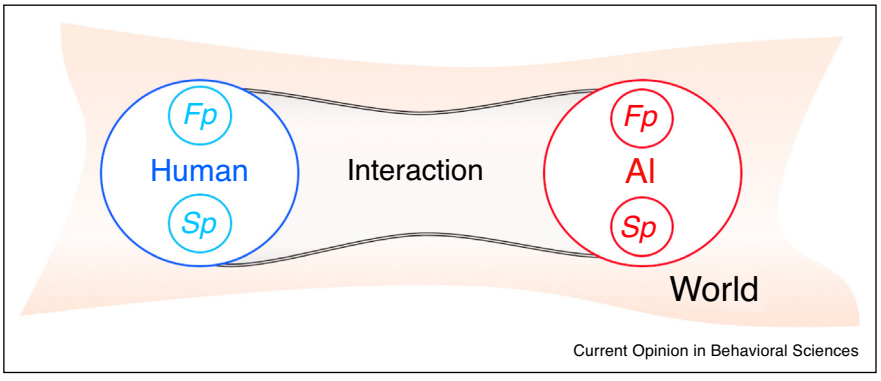
Synthesis of SoA theories

SoA is especially sensitive to any disruption in the smooth, harmonious flow of intentional actions [22] to expected sensory outcomes, that is, along the *intention–action–outcome* chain [23]. Disruptions along this chain can therefore inform loss of SoA. Thus, explaining SoA depends on identifying how it emerges from or gets disrupted in this chain.

The two most influential theories are the comparator model (CM) [24] (Figure 3a) and retrospective inference (RI) [25] (Figure 3b). Although originally a theory of motor learning and control, the CM's relevance to action awareness has been widely discussed. According to the CM, motor action is accompanied by a prediction of its outcome that is generated based on the copy of the motor command. This prediction is compared with the perceived sensory outcome: if there is no mismatch, then the outcome is registered as self-caused; otherwise, a disruption of SoA occurs. In contrast, RI rejects the strong involvement of sensorimotor mechanisms and posits that SoA results from cognitive sense-making processes. Specifically, SoA is inferred whenever there is congruence between the intended or logically expected outcome and the perceived outcome. While predicted outcomes emerge from sensorimotor processes before action (i.e. prospective), intended outcomes and thoughts are evaluated after action–outcome effects have been experienced (i.e. retrospective).

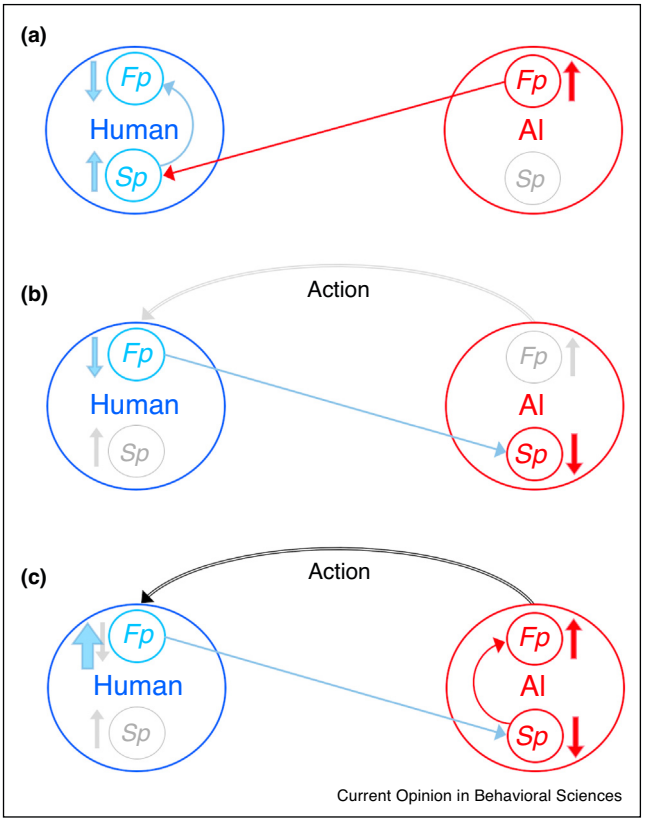
The multifactorial weighting model (MWM) [10] (Figure 3c) sought to find a compromise between the CM and RI [26], but more importantly to overcome the biological and explanatory disadvantages of the CM, for example, it could not explain agency during phantom

Figure 1



Coupling of sense of agency (SoA) in human–AI interaction with first-person (*Fp*) and second-person (*Sp*) perspectives of SoA. The AI or human possesses intrinsic SoA in *Fp*. In *Sp*, each has a model or mental representation of the other's *Fp*.

Figure 2



Dynamics of the first-person and second-person (*Fp* and *Sp*, respectively) perspectives of SoA in human–AI interaction. **(a)** The dilemma is that an AI simply demonstrating high SoA may reduce human SoA because the human perceives the AI to be controlling him/her. **(b)** The AI, upon perceiving the decrease in human SoA, must then carefully choose its next actions, or perhaps perform no action at all, to increase human SoA. **(c)** The AI, armed with an understanding of both human and its own SoA, can then better respond to improve, rather than erode, human SoA.

limb movements and thought insertions in schizophrenia (see Ref. [10] for details). According to MWM, SoA is separated into the feeling and judgment of agency (FoA and JoA, respectively). FoA is described as implicit, pre-reflective, low-level (involves sensorimotor processes), and non-conceptual, that is, it operates at the hem of consciousness [26]. Its components are similar to those of the CM. In contrast, JoA is described as explicit, interpretative and conceptual. It is based on FoA, but also on higher-order (cognitive) factors such as external contextual and social cues, as well as internal intentions and thoughts akin to the RI. Hence, FoA is computed mainly at the sensorimotor level without needing the concept of self and others, and JoA depends on logical inference and distinction of self and others [10]. According to the MWM, SoA arises from the constant weighting of agency cues according to their reliability.

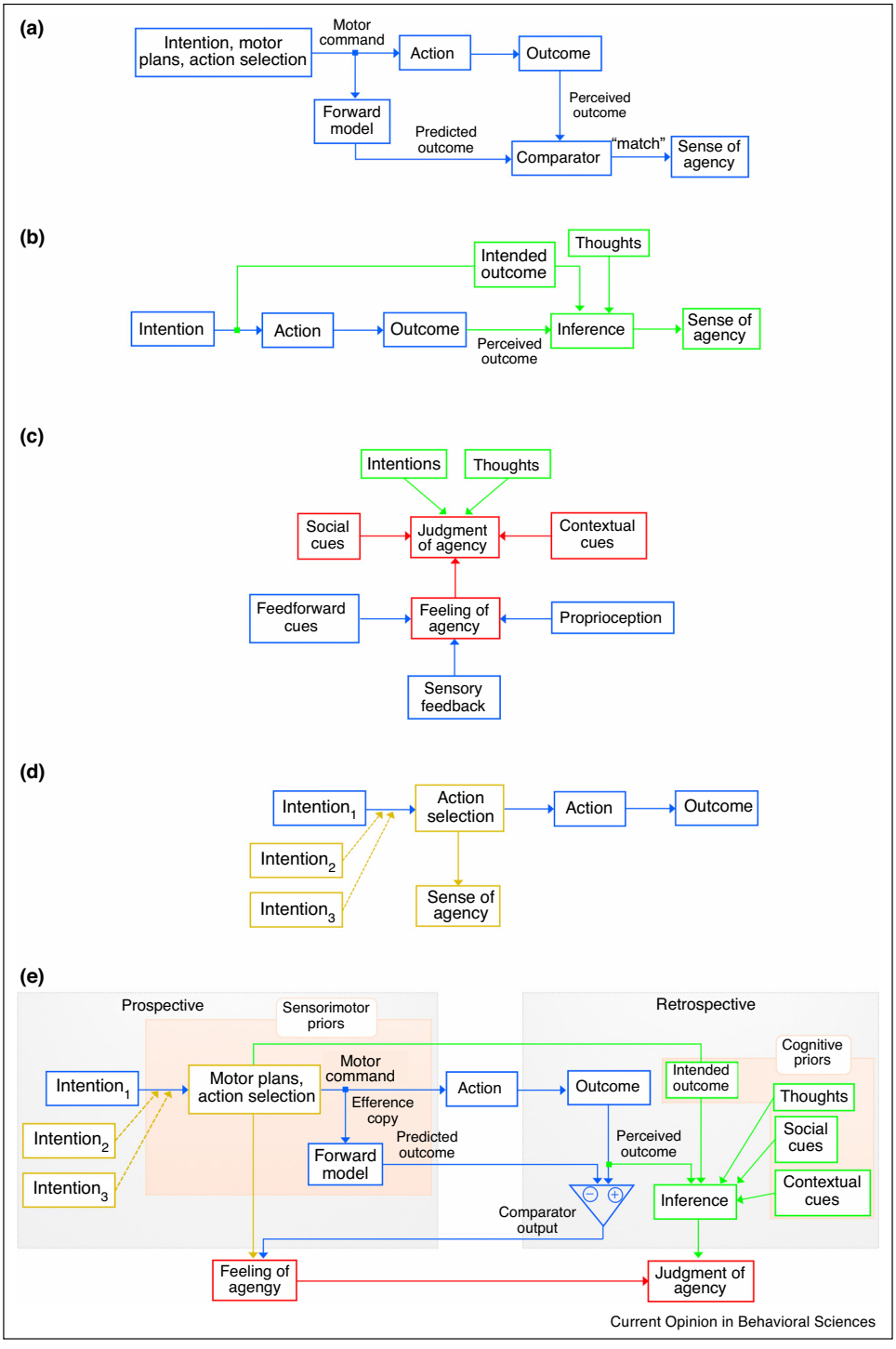
The major criticism to the MWM is that its initial account was ambiguous to how the brain assigns the weights and integrates agency cues in context-dependent situations. The Bayesian cue integration theory provides the mathematical foundation to realize this [27,28]. Here, SoA is formalized as a weighted combination of agency cues from different modalities, and each weight is a function of the accuracy in each modality. Moreover, and importantly, the weighting may be altered by prior knowledge or expectation, implemented as Bayesian priors, that may pertain to predictive sensorimotor signals or cognitive cues and action–outcome causal relations.

Lastly, another recent theory gaining traction is that a signal about the fluency of action selection can prospectively inform SoA [23] (Figure 3d). According to this theory, SoA is expected to increase when actions are freely chosen as opposed to actions that are instructed [29], coerced [30], or triggered involuntarily (e.g. by brain stimulation) [2••]. SoA is also expected to decrease when intentions that conflict with the about to be executed intention arise [23]. Others found SoA to increase with the number of alternative actions leading to the same outcome, and with the ability to select between actions with different foreseeable consequences [29].

We synthesize the theories above to show a coherent integration of prospective and retrospective components informing FoA or JoA (Figure 3e). On a sensorimotor level, the prospective account consists of *sensorimotor priors* [28]. These are agency cues in sensorimotor format that may be predictive (e.g. efference copy and internal predictions based on it) or they are not based on the predictability of the action outcome (e.g. conflict in intention and action fluency). Depending on the context and the environment, these internal signals can directly lead to FoA. On other occasions, internal predictions are compared to actual outcomes for retrospective FoA. At the higher level, the inference mechanism uses *cognitive priors* to inform JoA retrospectively. It is important to realize that these influential and recent theories were not derived from a vacuum, rather, they have been validated empirically to suggest their viability. Our synthesis can serve as 'roadmap' for an AI research to locate its position and from there rethink on how to gain more theoretical grounds to achieve stronger models of SoA.

We have explained thus far how SoA emerges or is disrupted along the intention–action–outcome chain. SoA, however, is phenomenologically thin, that is, we are most of the time only minimally aware of our agency when we act [1], which makes SoA hard to measure. This has prompted experimentalists to develop paradigms to measure it, which are either explicit or implicit [31]. Explicit measures directly ask subjects to report their agentic experience (e.g. whether the movement was theirs or not or how much they felt their action caused

Figure 3



Synthesis of influential and recent theories on the human science treatment of sense of agency (SoA). The colors highlight the theories. The comparator model **(a)** and retrospective inference **(b)** are most influential. The former hinges on matching perceived sensory outcomes to predicted outcomes that emerge from sensorimotor processes to inform SoA, and the latter relies on the congruence of perceived and intended outcomes, or related thoughts, when inferring SoA. The multifactorial weighting model **(c)** posits that SoA arises from the constant weighting of agency cues according to their reliability, and distinguishes between the feeling (FoA) and judgment of agency (JoA). More recently, a signal on fluency of action selection **(d)** has been theorized to prospectively inform SoA. We synthesize these theories, in **(e)**, to show the prospective and retrospective components informing agency. The prospective account consists of sensorimotor priors, which may directly inform FoA. At other times, internal predictions are compared to perceived outcomes. At the higher level, the inference mechanism uses cognitive priors to retrospectively inform JoA. Figures (a), (c), and (d) are based on Refs. [43], [44], and [45], respectively.

the outcome). Self-reports, however, are subject to cognitive biases and often blurred by unconscious thoughts. Implicit measures, in contrast, use perceptual differences between intentional and unintended action–outcome effects to measure SoA than directly ask about agentic experiences. The scientific consensus now is that intentional binding [32], that is, the perceived time of action and outcome shifted toward each other when SoA is high, is a robust implicit measure of SoA.

AI is poised to inform *Fp* SoA

The influence of the comparator model has reached developmental robotics, which allows for self-attribution of action–outcome effects in robots [8•,9]. A notable example is ego-noise attenuation, said to be one of the biggest and most unexplored problems in robot listening [9]. Ego-noise is the sound that the robot makes when moving. It has been shown that the incongruence of its proprioceptive information to the perceived ego-noise generates bigger prediction errors, which disturbs the robot with self-generated ego-noises. Self-organizing maps and multilayer perceptrons have been used to construct the forward model. Because of the CM's significance to motor learning and control, and SoA, it is understandable why it attracted focus in robotic SoA. As we explained above, however, the CM alone is insufficient to account for a full SoA.

When its SoA is disrupted, the AI would need to infer the cause and explain the nature of the disruption. Causal inference in AI research, albeit not focused on SoA, is not new and still well posited to revolutionize AI [33]. Pearl has recently postulated a three-layer causal hierarchy [33,34], and argues that it overcomes problems that machine learning has yet to hurdle, specifically, understanding and explaining cause–outcome relations. The causal hierarchy finds statistical relationships (correlation but not causation), intervention (causal effects), and counterfactuals (retrospective and explanatory). We posit that the last, which subsumes the other two [33], is imperative to *Fp* SoA. For example, if a boy missed a target with a toy gun but the target fell, he would have lower SoA if there were other kids shooting. But if he was alone, he would infer that his action caused the target to fall, and consequently have high SoA.

Studies that cast SoA as optimal Bayesian cue integration have yet to provide the mathematical elucidation. However, in a recent work [3•], the authors formalized SoA by drawing parallels from a Bayesian inference of the *ventriloquism effect* that estimates a common cause behind its multisensory integration. Their Bayesian model was able to concisely reproduce the intentional binding experiment. More importantly, the model explained the underlying computational mechanisms that drove this intentional binding effect. They theorized SoA as the *confidence in causal estimate* (*CCE*): it is high when action–outcome sensory signal is consistent with prior knowledge that the action causes the outcome, the causal belief is strong, and the action and outcome signals are reliable. This notion is consistent to SoA emerging from action–outcome consistency and from the reliability-dependent integration of different agency cues. Furthermore, they postulated *CCE* to fit the notion of FoA: it is a multimodal integration process that lies at the center of obtaining a Bayesian causality inference and does not attribute causality to any other agent. JoA, that is, judgment on causality, is made by comparing *CCE* with the confidence in acausal estimate, and SoA is attributed if *CCE* is higher.

Finally, action selection and intention conflict resolution are widely investigated topics in AI, but have not been investigated for the purpose of SoA in AI. Intention conflict resolution in AI has been demonstrated by both single (e.g. Refs. [35,36]) and multi-agent (e.g. Refs. [37,38]) reinforcement learners. Most of these learners would treat intention as an extra parameter, which means that conflict of intention can be detected if the agent is taking an action with a different intention parameter value. Action selection has been modeled as action planning using Markov decision processes, as well as model-based, model-free or combined reinforcement learning methods to learn policies [39,40]. The issue is how to quantify the *fluency* of action selection. This can be quantified relative to the number of alternative actions made available to efficiently optimize control. Another is to use the entropy of action selection probabilities, for example, if the typical softmax output layer is assumed for action selection probabilities, then fluency is related to the temperature parameter [41], that is, higher temperature means softer probability distribution and lower fluency.

Conclusion

Empirical evidence in human science demonstrates that the degradation of SoA characterizes certain psychiatric and neurological disorders [27,42,44]. These hinder the patients' ability to normally function mentally, emotionally or socially. Analogously but to a lesser extent, faced with an AI that is capable of volition, self-awareness, causal understanding and sense of social responsibility, an AI that is lacking SoA may be perceived as suboptimal and less sufficient in comparison since it lacks such capabilities. This should not be the case in the near future since, as we posited here, AI can readily draw knowledge from human science and use its advanced tools to realize stronger models of SoA. Perhaps, more robust studies on synthetic agency may later on inform better examinations of human SoA.

Conflict of interest statement

Nothing declared.

Acknowledgements

This study was supported by Brain/MINDS from AMED under Grant Number JP19dm020700 and JSPS KAKENHI Grant Number JP18H05432.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Gallagher S: **Multiple aspects in the sense of agency**. *New Ideas Psychol* 2012, **30**:5-31 http://dx.doi.org/10.1016/j.newideapsych.2010.03.003.

2. •• Haggard P: **Sense of agency in the human brain**. *Nat Rev Neurosci* 2017, **18**:196-207 http://dx.doi.org/10.1038/nrn.2017.14.

The psychological and neurocognitive bases of sense of agency and the progress that has been made are excellently elucidated in this paper.

3. • Legaspi R, Toyoizumi T: **A Bayesian Psychophysics Model of Sense of Agency**. *bioRxiv* 2018 http://dx.doi.org/10.1101/433888. 433888.

This work provides a strong mathematical elucidation of sense of agency, one that is missing in the current literature. It adapted a Bayesian inference model originally used to explain the ventriloquism effect to create a formal model that posits a precision-dependent causal sense of agency.

4. Barandian XE, Paolo E, Di Rohde M: **Defining agency: individuality, normativity, asymmetry, and spatio-temporality in action**. *Adapt Behav* 2009, **17**:367-386 http://dx.doi.org/10.1177/1059712309343819.

5. Chatila R, Renaudo E, Andries M, Chavez-Garcia RO, Luce-Vayrac P, Gottstein R, Alami R, Clodic A, Devin S, Girard B, Khamassi M: **Toward self-aware robots**. *Front Robot AI* 2018, **5**:88 http://dx.doi.org/10.3389/frobt.2018.00088.

6. Shneiderman B, Plaisant C: *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. edn 4. Reading, MA: Pearson Addison-Wesley; 2004.

7. Asada M, Hosoda K, Kuniyoshi Y, Ishiguro H, Inui T, Yoshikawa Y, Ogino M, Yoshida C: **Cognitive developmental robotics: a survey**. *IEEE Trans Auton Mental Dev* 2009, **1**:12-34 http://dx.doi.org/10.1109/TAMD.2009.2021702.

8. • Schillaci G, Hafner VV, Lara B: **Exploration behaviors, body representations, and simulation processes for the development of cognition in artificial agents**. *Front Robot AI* 2016, **3**:39 http://dx.doi.org/10.3389/frobt.2016.00039.

This paper provides a comprehensive survey of how sensorimotor mechanisms in developmental robots bridge sensorimotor representations and fundamental cognitive skills that include sense of agency. The focus is on the components of the comparator model of sense of agency.

9. Schillaci G, Ritter CN, Hafner VV, Lara B: **Body representations for robot ego-noise modelling and prediction. Towards the development of a sense of agency in artificial agents**. *Proceedings of the 15th International Conference on the Synthesis and Simulation of Living Systems (ALIFE XV)* 2016:390-397 http://dx.doi.org/10.7551/978-0-262-33936-0-ch065.

10. Synofzik M, Vosgerau G, Newen A: **Beyond the comparator model: a multifactorial two-step account of agency**. *Conscious Cogn* 2008, **17**:219-239 http://dx.doi.org/10.1016/j.concog.2007.03.010.

11. •• Synofzik M: **Comparators and weighting: neurocognitive accounts of agency**. In *The Sense of Agency*. Edited by Haggard P, Eitam B. Oxford University Press; 2015:289-306 http://dx.doi.org/10.1093/acprof:oso/9780190267278.003.0010.

This paper explains the limitations of the influential comparator model of sense of agency (references the widely cited version, i.e. Ref. [10]). It then explains why this limitation necessitated the shift in paradigm to what is now gaining attention, that is, the optimal Bayesian cue integration model of sense of agency.

12. Dennett DC: *The Intentional Stance*. Cambridge, MA: The MIT Press; 1987.

13. Złotowski J, Proudfoot D, Yogeeswaran K, Bartneck C: **Anthropomorphism: opportunities and challenges in human-robot interaction**. *Int J Soc Robot* 2015, **7**:347-360 http://dx.doi.org/10.1007/s12369-014-0267-6.

14. Wang Z, Mülling K, Deisenroth MP, Amor HB, Vogt D, Schölkopf B, Peters J: **Probabilistic movement modeling for intention inference in human-robot interaction**. *Int J Robot Res* 2013, **32**:841-858 http://dx.doi.org/10.1177/0278364913478447.

15. Ramirez-Amaro K, Beetz M, Cheng G: **Understanding the intention of human activities through semantic perception: observation, understanding and execution on a humanoid robot**. *Adv Robot* 2015, **29**:345-362 http://dx.doi.org/10.1080/01691864.2014.1003096.

16. Vrigkas M, Nikou C, Kakadiaris IA: **A review of human activity recognition methods**. *Front Robot AI* 2015, **2**:28 http://dx.doi.org/10.3389/frobt.2015.00028.

17. Buchsbaum D, Griffiths TL, Plunkett D, Gopnik A, Baldwin D: **Inferring action structure and causal relationships in continuous sequences of human action**. *Cogn Psychol* 2015, **76**:30-77 http://dx.doi.org/10.1016/j.cogpsych.2014.10.001.

18. Berberian B, Sarrazin JC, Le Blaye P, Haggard P: **Automation technology and sense of control: a window on human agency**. *PLoS One* 2012, **7**:e34075 http://dx.doi.org/10.1371/journal.pone.0034075.

19. Limerick H, Coyle D, Moore JW: **The experience of agency in human-computer interactions: a review**. *Front Hum Neurosci* 2014, **8**:643 http://dx.doi.org/10.3389/fnhum.2014.00643.

20. Tao J, Tan T: **Affective computing: a review**. In *Affective Computing and Intelligent Interaction.* Edited by Tao J, Tan T, Picard RW. Berlin, Heilderberg: Springer 3784; 2005:981-995 http://dx.doi.org/10.1007/11573548_125. Lecture Notes in Computer Science.

21. Yonck R: *Heart of the Machine: Our Future in a World of Artificial Emotional Intelligence*. New York, USA: Arcade Publishing; 2017.

22. Brass M, Haggard P: **The what, when, whether model of intentional action**. *The Neuroscientist* 2008, **14**:319-325 http://dx.doi.org/10.1177/1073858408317417.

23. Chambon V, Sidarus N, Haggard P: **From action intentions to action effects: how does the sense of agency come about?** *Front Hum Neurosci* 2014, **8**:320 http://dx.doi.org/10.3389/fnhum.2014.00320.

24. Frith CD, Blakemore SJ, Wolpert DM: **Abnormalities in the awareness and control of action**. *Philos Trans R Soc Lond B Biol Sci* 2000, **355**:1771-1788 http://dx.doi.org/10.1098/rstb.2000.0734.

25. Wegner DM, Wheatley T: **Apparent mental causation: sources of the experience of will**. *Am Psychol* 1999, **54**:480-492 http://dx.doi.org/10.1037/0003-066X.54.7.480.

26. Braun N, Debener S, Spychala N, Bongartz E, Sörös P, Müller HHO, Philipsen A: **The sense of agency and ownership: a review**. *Front Psychol* 2018, **9**:535 http://dx.doi.org/10.3389/fpsyg.2018.00535.

27. Moore JW, Fletcher PC: **Sense of agency in health and disease: a review of cue integration processes**. *Conscious Cogn* 2012, **21**:59-68 http://dx.doi.org/10.1016/j.concog.2011.08.010.

28. Synofzik M, Vosgerau G, Voss M: **The experience of agency: an interplay between prediction and postdiction**. *Front Psychol* 2013, **4**:127 http://dx.doi.org/10.3389/fpsyg.2013.00127.

29. Barlas Z, Obhi SS: **Freedom, choice, and the sense of agency**. *Front Hum Neurosci* 2013, **7**:514 http://dx.doi.org/10.3389/fnhum.2013.00514.

30. Caspar EA, Christensen JF, Cleeremans A, Haggard P: **Coercion changes the sense of agency in the human brain**. *Curr Biol* 2016, **26**:585-592 http://dx.doi.org/10.1016/j.cub.2015.12.067.

31. Dewey JA, Knoblich G: **Do implicit and explicit measures of sense of agency measure the same thing?** *PLoS One* 2014, **9**:e110118 http://dx.doi.org/10.1371/journal.pone.0110118.

32. Moore JW, Obhi SS: **Intentional binding and the sense of agency: a review**. *Conscious Cogn* 2012, **21**:546-561 http://dx.doi.org/10.1016/j.concog.2011.12.002.

33. Pearl J: **The seven tools of causal inference with reflections on machine learning**. *Commun ACM* 2019, **62**:54-60 http://dx.doi.org/10.1145/3241036.

34. Pearl J: *Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution*. 2018 . arXiv 1801.04016 *https://arxiv.org/pdf/1801.04016.pdf*.

35. Kulkarni TD, Narasimhan KR, Saeedi A, Tenenbaum JB: **Hierarchical deep reinforcement learning: integrating temporal abstraction and intrinsic motivation**. *Adv Neural Inf Process Syst* 2016, **29**:3675-3683 https://papers.nips.cc/paper/6233-hierarchical-deep-reinforcement-learning-integrating-temporal-abstraction-and-intrinsic-motivation.pdf.

36. Zhu Y, Mottaghi R, Kolve E, Lim JJ, Gupta A, Fei-Fei L, Farhadi A: **Target-driven visual navigation in indoor scenes using deep reinforcement learning**. *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)* 2017:3357-3364 https://doi.org/10.1109/ICRA.2017.7989381.

37. Foerster JN, Assael YM, de Freitas N, Whiteson S: **Learning to communicate with deep multi-agent reinforcement learning**. *Adv Neural Inf Process Syst* 2016, **29** https://papers.nips.cc/paper/6042-learning-to-communicate-with-deep-multi-agent-reinforcement-learning.pdf.

38. Lowe R, Wu Y, Tamar A, Harb J, Abbeel P, Mordatch I: **Multi-agent actor-critic for mixed cooperative-competitive environments**. *Adv Neural Inf Process Syst* 2017, **30**:6379-6390 https://papers.nips.cc/paper/7217-multi-agent-actor-critic-for-mixed-cooperative-competitive-environments.pdf.

39. Kober J, Bagnell JA, Peters J: **Reinforcement learning in robotics: a survey**. *Int J Robot Res* 2013, **32**:1238-1274 http://dx.doi.org/10.1177/0278364913495721.

40. Renaudo E, Girard B, Chatila R, Khamassi M: **Respective advantages and disadvantages of model-based and model-free reinforcement learning in a robotics neuro-inspired cognitive architecture**. *Procedia Comput Sci* 2015, **71**:178-184 http://dx.doi.org/10.1016/j.procs.2015.12.194.

41. Hinton G, Vinyals O, Dean J: *Distilling the Knowledge in a Neural Network*. . arXiv 1503.02531 2015 *https://arxiv.org/abs/1503.02531*.

42. Rowe JB, Wolpe N: **Disorders of volition from neurological disease: altered awareness of action in neurological disorders**. In *The Sense of Agency*. Edited by Haggard P, Eitam B. Oxford University Press; 2015:389-414 http://dx.doi.org/10.1093/acprof:oso/9780190267278.003.0018.

43. Morsella E, Berger CC, Krieger SC: **Cognitive and neural components of the phenomenology of agency**. *Neurocase* 2011, **17**:209-230 http://dx.doi.org/10.1080/13554794.2010.504727.

44. Robinson JD, Wagner NF, Northoff G: **Is the sense of agency in schizophrenia influenced by resting-state variation in self-referential regions of the brain?** *Schizophr Bull* 2016, **42**:270-276 http://dx.doi.org/10.1093/schbul/sbv102.

45. Sidarus N, Haggard P: **Difficult action decisions reduce the sense of agency: a study using the Eriksen flanker task**. *Acta Psychol* 2016, **166**:1-11 http://dx.doi.org/10.1016/j.actpsy.2016.03.003.