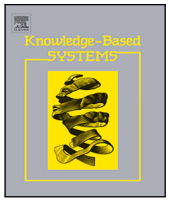
Knowledge-Based Systems 286 (2024) 111298

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



The sense of agency in human–AI interactions

Roberto Legaspi [a,*], Wenzhen Xu [a,d], Tatsuya Konishi [a], Shinya Wada [a], Nao Kobayashi [b], Yasushi Naruse [c], Yuichi Ishikawa [a,e]

[a] Human-Centered AI Laboratories, KDDI Research, Fujimino, Japan
[b] Life Science Laboratories, KDDI Research, Fujimino, Japan
[c] Center for Information and Neural Networks (CiNet), National Institute of Information and Communications Technology, Kobe, Japan
[d] Business Administration of Hitotsubashi University, Japan [1]
[e] Mitsubishi Research Institute Inc., Japan[1]

ARTICLE INFO

Keywords:
Sense of agency
Machine agency
Human–AI interaction

ABSTRACT

Sense of agency (SoA) is the perceived control over one's actions and their consequences, and through this one feels responsible for the consequent outcomes in the world. We analyze the far-reaching implications of a two-pronged knowledge on SoA and its impact on human–AI interactions. We argue that although there are interesting research efforts for an AI to inherently possess SoA, they are still sparse, constrained in scope and present unclear immediate benefit to the design of AI-enabled systems. We also argue that the knowledge on how human SoA is affected by an AI that is perceived to possess a sense of control presents more immediate benefit to AI, in particular, to eliciting positive human attitudes toward AI. Third, and lastly, we argue that research efforts for an AI to adapt to the dynamic changes of human SoA are practically non-existent primarily due to the difficulty of modeling, inferring and adaptively responding to human SoA in complex natural settings. We proceed by first delving deep into the influential and recent theoretical underpinnings of SoA, and discuss its conceptual reach in different disciplines and how it is applied in real-world research. We organize a substantial part of our paper to put forward and elucidate our three argumentative points while supported by evidence in the literature.

1. Introduction

The *sense of agency* (SoA) is the subjective experience or judgment of control over one's own actions that caused changes in the world, and thus one feels responsible for the consequent outcomes (e.g., "I was the one who made the ball bounce back. A neat trick indeed!") [1–6]. Research on SoA once existed in obscurity, pent to a group of scholars centered on epistemic questions regarding free will and consciousness [2]. The experience that we freely choose the actions to perform, and the sense that we claim authorship of those actions are central aspects of our consciousness [7]. We can anticipate a strong SoA with volitional versus instructed [8] or coerced [9] actions, when there are different actions to choose from all leading to the same intended outcome, with increased capacity from choosing among actions that have varying anticipated consequences [8], and when there are no conflicting intentions [10]. Researchers also discovered that SoA disturbances are a crucial component of psychosis and are typical of schizophrenia spectrum illnesses [4,5,11–18]. It has been posited that aberrant prediction mechanisms may induce inaccurate self-referencing experiences, which result in a failure to recognize specific cognitions as self-generated or self-caused [18–20].

The reach of SoA research, however, has gone beyond knowledge related to consciousness, free will and mental pathology. For instance, evidence suggest a strong overlap between SoA and HCI [21–25]), as the latter has long acknowledged that preserving user's sense of being in control is essential when engaging with technology [26]. Such as in terms of input modalities (see also [27,28]), system feedback mechanism and computer assistance, among others, with the overarching consideration of SoA research informing HCI and vice versa [22]. Another is that research on human-machine interaction (HMI) found that increasing machine automation weakens SoA and the perception of action-outcome causal effects [23,25,29–33]. Finally, when it comes to the SoA construct being studied in human and AI interactions, investigations remain scarce and theoretical. Most if not all current works fall within robots that simulate a human in early cognitive development [34–37], for instance, a robot that uses sensorimotor predictive processes to distinguish actions initiated by itself from those

* Corresponding author.
E-mail address: ro-legaspi@kddi-research.jp (R. Legaspi).
[1] At the time of publication.

https://doi.org/10.1016/j.knosys.2023.111298
Received 30 September 2022; Received in revised form 5 December 2023; Accepted 12 December 2023
Available online 17 January 2024

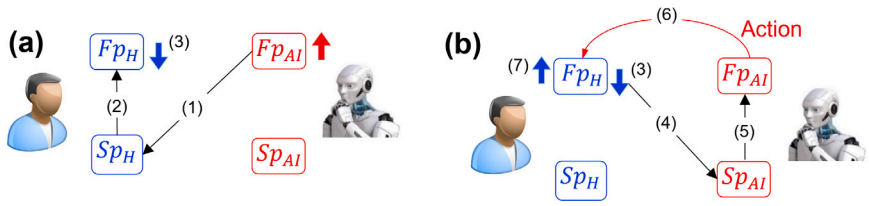


Fig. 1. Dynamics of the first and second-person perspectives (Fp and Sp, respectively) of SoA in human–AI interactions by Legaspi et al. [43]. (The numbers in parentheses indicate the sequence of events in the figure.) (a) The problem arises when an AI demonstrating strong Fp_{AI} (synthetic SoA) weakens the Fp_H (human SoA) because the human judges via its Sp_H the AI as having greater control of the interaction. (b) The AI, realizing through its Sp_{AI} that Fp_H has weakened, must then cautiously decide its next actions for the human to regain or increase its Fp_H. Thus, by making sense of Fp_H the AI can then reasonably and empathically adapt to, instead of disregard, human SoA.

of other agents around it, consequently enhancing its sensorimotor and cognitive skills [38–42].

We begin, however, on a recent paper [43] that proposed a two-person view of SoA for human–AI interactions (Fig. 1). In the first-person perspective (Fp), the AI, and not just the human (denoted hereafter with "H"), intrinsically has SoA. In the second-person (Sp), the human and AI evaluate the other's SoA using a constructed model of their counterpart's Fp. The critical issue is when a strong Fp_{AI} negatively influences Sp_H, which consequently causes the Fp_H to weaken (Fig. 1(a)). In other words, when an AI demonstrates a strong SoA, it weakens the human SoA because the human judges the AI as having greater control of the interaction. This can be true in human–human interactions, in which the strength of the partner's self-agency depends on the role and amount of control each partner has in the social interaction [9,44–47]. However, the AI being cognizant of the weakening Fp_H will then watchfully plan and select its succeeding actions, or perhaps resort to inaction, if only to relinquish control back to the human in order to regain or improve the Fp_H (Fig. 1(b)).

Illustrative case in point: a middle-aged woman was hit by an Uber self-driving car as she wheeled a bicycle across a road in Tempe, Arizona in 2018. The vehicle's recognition systems failed to identify her and her bicycle as imminent collision danger, and so the car did not automatically stop. However, authorities concluded that human error was mostly to blame since the back-up driver could have taken control of the vehicle in case of an emergency, but dash-cam footage showed her streaming a TV show episode and taking her eyes off the road for several seconds immediately before the hit. We surmise the two-pronged SoA perspective could be at play in this tragic incident. The driver could have resumed manual control at any time (indicating that Fp_H is not constrained), but became overly reliant (diminished Fp_H) on the car's automated control, and perceived (through her Sp_H) a high car agency, albeit the car did not physically possess an actual Fp_{AI}. Consequently, the driver disengaged and her intervention to regain manual control became delayed (severely diminished Fp_H), which led to the fatality. Had the self-driving car been cognizant (via its Sp_{AI}) of the driver's low SoA, it could have relinquished control back to the driver to sustain safety or probed the driver's engagement level (see [30]) and trigger an SoA recovery when critical. Indeed, an automated vehicle with image or physiology recognition capabilities can be made aware of its operator being inattentive or disengaged from the external environment, which should have been enough condition for it to pulse the driver's sense of awareness. A recent work [48], however, demonstrated that even when vehicle monitoring successfully prompted operators to keep the wheels in their hands and their eyes on the road, they still disengaged from driving and were unable to avoid crashing the vehicle. What we surmise is that reduced SoA greatly influences the operator's attention allocation and interventional responses. When operators, who have assumed supervisory roles since the physical translation of their intentions has now been carried out by the machine and are no longer actively involved in the task [31], become overconfident in the system's capabilities to successfully complete their task through automation [49], this materializes in reduced SoA [23,25,29–31]. This reduced SoA then leads to disengagement since the machine has appeared (or been perceived) to be an intentional entity possessing agency attributes, in this case, being able to attend to the task (e.g., driving), which the operator can now afford not to do (see [30,50,51]). Regaining SoA is a viable solution to the operator's impaired performance [49].

We reckon Legaspi et al.'s two-person perspective of SoA deserves further attention because of the compelling nature of its hypotheses in light of enhancing knowledge in human–AI interaction. Legaspi and colleagues advocated for the creation of what they call *synthetic agency*, i.e., SoA possessed by an AI. They contend that the methods required to construct a synthetic agency are now available in AI research and that the field is poised to inform strong models of synthetic agency. For them to aim for such a goal seems understandable, as they are coming from their neuro-cognitive, behavioral, and theoretical AI perspectives. However, as we farmed our way through the SoA, HCI, HMI and HRI literature, we realized that much has yet to be discussed and debated in terms of its feasibility and viability. From a practical, application vantage point, understanding how human SoA is influenced by an AI that is perceived to possess (not actually possessing) agency presents immediate benefits to the design of AI systems, particularly from eliciting positive human attitudes toward AI (such as trustworthiness). Moreover, an AI responding adaptively to its understanding of the dynamic changes in human SoA to help improve it is less than nascent but worth exploring. Against this backdrop, we argue for the following points:

- Research efforts to actualize Fp_{AI}, i.e., an AI with innate SoA, remain sparse, constrained in scope and present unclear immediate benefit.
- Research on $Fp_{AI} \rightarrow Sp_H \rightarrow Fp_H$ (Fig. 1(a)), i.e., the understanding of how human SoA is affected by an AI that is perceived to possess agency, presents more immediate benefit to the design of AI-enabled systems, particularly from eliciting positive human attitudes toward AI.
- Research on $Fp_H \rightarrow Sp_{AI} \rightarrow Fp_{AI} \rightarrow Action$ (Fig. 1(b)), i.e., an AI responding adaptively to its understanding of the dynamic changes in human SoA to consequently improve it, albeit compelling, is practically non-existent, and therefore needs to be pursued.

We proceed by elucidating the influential theoretical underpinnings of the compelling nature of SoA. We then organize the rest of our paper to substantially argue our three positions above, supported by evidence in the literature, and equally important, put forward for consideration our computational framework as the first step toward going over the third argument.

2. Theoretical underpinnings of the sense of agency

The influential and recent theories from the cognitive, neuro and behavioral science treatments of SoA posit that it emerges from *prospective* (before the action is made) and *retrospective* (after the sensory outcome is perceived) processes [52,53]. While the former is based on high-level cognitive monitoring of how actions are selected, i.e., linking intentions to action [10], the latter relies on predictive and postdictive

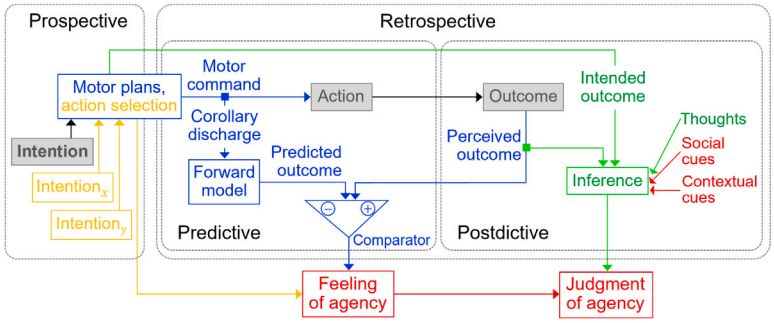


Fig. 2. Synthesis of the prospective and retrospective accounts of SoA per Legaspi et al. [43]. The comparator model (CM) [57] (in blue) and retrospective inference (RI) [58] (green) are classical and prominent. While the CM informs SoA by matching the perceived and predicted outcomes, RI depends on the consilience between the outcome that is perceived and the intended outcome or the accompanying related thoughts. Fluency of action selection (in gold) is also theorized to prospectively inform SoA that is expected to diminish when conflicting intentions are present [10]. The multifactorial weighting model (in red) suggests that SoA emerges from weighting and integrating various agency cues and discerns between the low-level feeling and high-level judgment of agency [59].

signals [54,55]. While most research separately treat the contributions of the prospective and retrospective accounts to eliciting SoA, they may, however, interact in certain contexts in which pre-action execution cues are combined with retrospective information on action consequences [56].

Majority of investigations have focused on the importance of the retrospective account in which the emergence of SoA is primarily interpreted as either the result of unerring direct access to sensorimotor priors preceding action execution for setting in place an agency experience (in the predictive), as in the case of the well-known *comparator model* (CM) [57,60] (Fig. 2 in blue), or the product of a fallible after the fact inference following prior beliefs and expectations, social interactions and context recognition during and after the action has occurred (in the postdictive), as in the case of *retrospective inference* (RI) [58,61,62] (Fig. 2 in green). The CM explains that if the prediction of the action's consequence that is produced from the corollary discharge that accompanies the motor action matches the recognized actual action outcome, then the outcome is attributed to the self's action; otherwise, the SoA is disrupted. In contrast to CM, the RI suggests SoA results from cognitive sense-making processes rather than through sensorimotor mechanisms.

The multifactorial weighting model (MWM) [59] (Fig. 2 in red) strikes as middle ground of the CM and RI [63], and suggests that SoA can be described as either implicit, pre-reflective, and non-conceptual, referred to as the feeling of agency that is akin to CM; or explicit, interpretative and conceptual, called judgment of agency, which is akin to RI. The MWM posits the brain integrates while weighting the relative influence of various agency cues based on their reliability. Thus, MWM has been formalized as Bayesian: SoA is a weighted sum of the multimodal agency signals wherein the weights are relative to the accuracy of each modality. Further, the values of the weights may change depending on the Bayesian priors that may refer to perceptual, mental or contextual cues.

The recent and interesting optimal Bayesian cue integration based formalism of Legaspi and Toyoizumi [64] postulates SoA as the *confidence in causal estimate* (CCE), that is,

$$CCE = \frac{\sigma_{tot}}{2\pi\sigma_A\sigma_O\sigma_{AO}} p_{causal}, \tag{1}$$

where σ_A, σ_O and σ_{AO} denote action, outcome and relative uncertainties, respectively, $\sigma^2_{tot} \equiv \sigma^2_A + \sigma^2_O + \sigma^2_{AO}$, and p_{causal} is the probability of the action causing the outcome. This model was able to replicate two key empirical settings of SoA and found CCE to correlate with SoA: CCE is high, medium and low for voluntary, involuntary and no intentional action conditions, respectively (consistent with [65]), and that CCE decreases with increasing uncertainty when perceiving the outcome (consistent with [66]). This finding may be viewed as unconventional, i.e., this model now explains SoA arises whether the actions were intended or not, which therefore suggests intentionality is not mandatory, as long as the confidence in the action causing the outcome is high. Furthermore, with reliable sensory cues, SoA can arise from unintentional actions. Thus, SoA is formalized in [64] as precision-dependent causal agency.

The prospective account of SoA is more recent and is based on the metacognitive monitoring of action selection processes [10,52,53,56] (Fig. 2 in gold). This signal on the fluency of choosing an action makes for a link between intentions and actions, and is provided before the action is executed, thus, also prospectively informs SoA [53]. Smooth or easy action selection yields stronger SoA than when actions are dysfluent or difficult, e.g., when we are forced to do something that goes against what we intend [9]. Thus, SoA diminishes when we perceive conflicting intentions [10]. Furthermore, an increase in SoA is expected when there are several actions to choose from and with increased ability to choose between actions with different predictable outcomes [8].

The output of the prospective and retrospective processes is a representation of self-agency, but the self need not be aware of it. This representation is most of the time unconscious, but since it is fashioned with every action, it should be present each time the self seeks it, or more generally when it is expected to emerge consciously [67]. In other words, we may ask for example, "Is this perceptual experience had by me?" If there is the uninterrupted, congruous stream of intentional actions to predicted sensory outcomes, and therefore nothing informs of any discordance in the intention-action-outcome chain (Fig. 2 in gray blocks) that disrupts our SoA, then we do not normally go through a reflective process to judge the experience of SoA to be our own; but if something contrary to what we expect occurs, then we become conscious to ask if the self is the subject of the experience [1]. In this instance, SoA is conceived as an in-the-moment but fleeting instance of self-awareness [68]. It is this introspection that leads to, among other things, thorny, controversial, and pervasive issues confronting AI research. This sets the backdrop of our first argument.

3. The first argument

Research efforts to actualize Fp_{AI}, i.e., an AI with innate SoA, remain sparse, constrained in scope and present unclear immediate benefit.

We begin our arguments by answering a fundamental question on the overlap of SoA and autonomous intelligent artifacts, i.e., given the theoretical underpinnings of how the representation of self-agency is elicited or disrupted (per Section 2), how much Fp_H is actually accounted for in current AI toward realizing an actual Fp_{AI}? Most research on actualizing Fp_{AI} are developing robots that are similar to a human toddler who is in the early stages of its cognitive development. Here a robot sets itself apart from other entities (e.g., objects, humans, and other robots) through sensorimotor predictive processes (e.g., in [38,40,42]). For instance, Schillaci et al. [38,39] suggest that

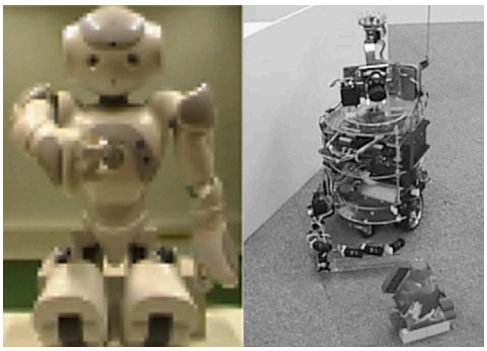


Fig. 3. Developmental robotics to realizing Fp_{AI}. The left image shows the Nao humanoid robot used by Schillaci et al. in [39] that performs sensorimotor predictions, and the right shows the vision-based mobile robot used by Tani in [35] in the late 1990s to help interpret the notion of "self" from a dynamical systems perspective.

providing agents the embodiment and enabling them with sensorimotor predictions is a promising direction toward the development of SoA in robotics. They adopted for their models dynamic self-organizing maps together with a Hebbian paradigm for the online learning of their humanoid robot (left image in Fig. 3). Their experiments revealed that sensory attenuation is more salient when the robot is the one who made the action, but worse when the mismatch of proprioceptive and motor information generates larger prediction errors. Sensory attenuation describes a phenomenon associated with normal human movements in which there is varying perception of the same sensory input contingent on the action being self or exogenously generated [69]: a self-generated stimulus is associated with the decrement in the perceived intensity of that stimulus (e.g., while one cannot tickle oneself, one can be tickled by others [70]). For Schillaci et al. this makes for an element of "surprise" for the robot that nudges it to recognize which action is generated by itself.

We can argue from the above that current attempts toward realizing Fp_{AI} mostly cover the predictive accounts at the sensorimotor levels and limited to developmental robotic agents, with no consideration yet to AI that is virtually embodied or as networks of embedded, ambient and wearable sensors and actuators that comprise the AI's embodiment (to be discussed in length in the next section). More important, however, is that it remains to be seen how the other processes – postdictive and its integration with the predictive account, and prospective and its integration with the retrospective account – can be modeled in an embodied AI for a more robust Fp_{AI}. Furthermore, we argue that it is not the case in the works above that a robot is aware that it elicits a representation of self-agency nor that the representation is elicited in the unconscious. In fact, the way to measure this in an embodied AI remains to be conceived.

A seeming step moving forward is CCE (Eq. (1)) that allows to quantitatively replicate the intentional binding (IB) effect in two well-known human experiments on SoA. The common consent right now in the SoA community is that the IB effect [71] is a robust assessment of implicit SoA. How we experience our self-agency can be *implicit*, i.e., we do not reflect nor conceptualize the feeling of being an agent [13,59] and this feeling emerges only within our subconscious [13], until we experience the urge to make the *explicit* judgment, i.e., one that is subjected to our interpretation and conceived notions of our agency [59]. Thus, the precision-dependent causal agency as computed by CCE in which SoA may arise as long as the sensory stimuli are perceived accurately, i.e., an embodied agent that is equipped with state of the art sensors that can perceive in sharp precision the stimuli may be interpreted as demonstrating some degree of implicit SoA.

But even if we reach the stage that AI can fully account for the prospective and retrospective underpinnings of SoA, we shall still need to establish whether an embodied AI can undergo a subjective experience. For instance, Tani [35] interpreted the "self" from a dynamical systems point of view through a robot (right image of Fig. 3) that integrated bottom-up information, i.e., derived from sensorimotor processes that represent environmental contexts, and top-down information, i.e., a predictive model of the environment. The robotic processes would remain uninterrupted so long as a good relative match is present with the two types of information. In a moment of incoherence, however, the robot's attention is directed to "be aware of" the mismatch to be resolved. For Tani, this is the robotic equivalent of "self-consciousness". However, there was no measure that existed to validate that Tani's robot is actually self-conscious.

More than two decades after Tani's work and despite major progress in robotics, we are still in search of an authentic theory on the underlying processes that could yield robotics and AI that are self-conscious [72] and aware [73]. Indeed, interesting and controversial discussions have been going on about ascribing a sense of self [74], consciousness [72], awareness [73], and even legal personhood to robots [75,76] (e.g., citizenship was granted to the humanoid robot Sophia in 2017), among others. However, these are still rightfully necessary discussions, but we are nowhere near systems achieving general intelligence. Similarly, we argue that we are still far from actualizing, let alone validating Fp_{AI} – not even fully understanding why, or if at all, we need or desire our AI to have Fp_{AI}. Anecdotally, after presenting the idea in a workshop[2] of theoretical and applied researchers on SoA, the first author of this article was asked, "But do we really need an AI to have SoA?" In HCI, perceived control is the belief that one's actions significantly change outcomes by causing desired results and preventing undesirable ones [77]. With the proven weightful influence of the feeling that we are in control of the way we think, feel, and behave, it should be expected that an emerging technology that would seem to undermine our desire to be in control may be viewed as unpleasant. Reactance theory even suggests that our loss of control over a situation may result in our aggression toward that which restricts our freedom. Perhaps the question should not only be a matter of need but also want: "Do we want an AI to have SoA?" We have yet to see this kind of discussion in the public arena, but it is one that we delve into in the next section. We can then conclude here that although there are efforts to achieve Fp_{AI}, they are still few and severely limited in their coverage of the SoA construct with no clear immediate social benefit nor unequivocal consensus to wanting to actualize Fp_{AI}.

4. The second argument

Research on $Fp_{AI} \rightarrow Sp_H \rightarrow Fp_H$, i.e., the understanding of how human SoA is affected by an AI that is perceived to possess agency, presents more immediate benefit to the design of AI-enabled systems, particularly from eliciting positive human attitudes toward AI.

Technology has greatly evolved via the creation of more sophisticated AI-enabled systems that accompany (e.g., home service robots [78], affective social agents [79,80] and shopping mall ushers [81]), aid (e.g., caretakers [82,83]) or substitute humans (e.g., in factories [84]) in many daily tasks. As more of such systems are developed for consumer use, it is crucial that their designs favor user acceptance, especially since they come with high price for both manufacturer and consumer. Second, investigations of humans interacting with artificial systems can benefit from a systematic examination of human behavior processes involved in their social interactions using methods with high degree of ecological validity and well-defined experimental control (see [85]). Related, a question that currently comes into view is how

[2] Psychological and Agency Workshop 2019 at https://sites.google.com/view/agency2019/home.

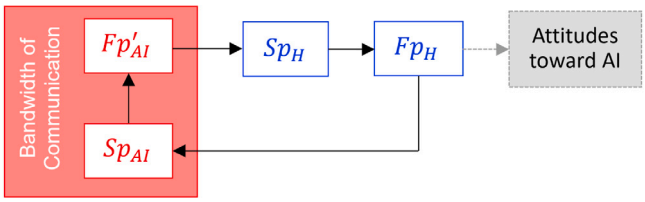


Fig. 4. Conceptual model. What goes into Sp_H need not even be the product of a real Fp_{AI}, but only a human-perceived sense of control in the AI (denoted by Fp'_{AI}), for it to affect Fp_H. Our conceptual model posits that the effect of Sp_H on one's attitude toward AI is mediated by Fp_H, i.e., as long as Fp_H is high, attitudes toward AI is positive even if the AI is perceived to have strong agency. Further, our conceptual model also posits that the greater the bandwidth of communication the AI possesses the greater the attribution of control and causal agency is given to the AI.

human SoA could be revised during joint interactions with technological systems versus isolated or concerted actions with other humans (pertinent reviews in [22,86,87]). Arguably, addressing these problems might provide new perspectives on the foundations of SoA and potentially impact how autonomous intelligent systems are developed in order to improve their effectiveness and integration into human civilization [87].

In this section, we are particularly concerned with human SoA being altered by an AI's capacity for control and causal action in joint agency. We adapt into our conceptual model (Fig. 4) the two-person SoA perspective that now explains Fp_H is affected even just by the perception that the AI has innate SoA. In other words, the Fp_{AI} is not real, but Sp_H thinks it is. We denote this human-perceived SoA in AI as Fp'_{AI}. This is not an empty conjecture as there is strong evidence to suggest that an *actual* Fp_{AI} is not a necessary condition to influence Fp_H. We humans do not shy away from attributing cognitive states or behavioral features to artificial systems even though we are cognizant that they do not authentically have such traits as long as they display even slight human characteristics or act in accordance to a social interaction rule (e.g., even earlier computers were already perceived as socially behaving agents [88,89]). This comes under the rubrics of intentional stance [90] and anthropomorphism [91]. For instance, the recent work in [92] demonstrated that artificial agents can cause in humans a vicarious SoA, and in a real physical setting shared by human and embodied artificial partner, intentionality attribution and anthropomorphic shape are important factors in the induction of vicarious SoA [92–94]. We are in an era where the sociability of artificial systems is no longer an oddity since they display human-like intelligence and in several instances surpasses human abilities. This renders such systems even more soliciting of human attribution and social response. We can therefore argue that, overall, researchers broadly agree that current artificial systems do not and need not inherently possess SoA. Thus, per our conceptual model, what goes into Sp_H need not even be products of a genuine Fp_{AI}, just an Fp'_{AI} is sufficient to have an eventual effect on Fp_H.

The second important aspect of our conceptual model is that it suggests the effect of the bandwidth of communication: the greater the bandwidth the AI has, the more computational power Fp'_{AI} and Sp_{AI} have, which may significantly impact the Sp_H. This bandwidth of communication increases with more natural kinds of capabilities afforded to the AI, such as, inter alia, perception, locomotion, object manipulation, speech, gesture, facial expression and display of affective states [95]. Historically, the notion of agency in AI has been catalyzed by the behavior-based robotics movement, which were mostly dynamical systems based approaches [96] that saw the adaptive real-world capacities of physically situated and embodied robots exceeded those of disembodied software. Studies show that supposedly social artifical agents are even more believable when embodied [97]. When we tried to place examples of embodied social agents on Milgram and Kishino's reality-virtuality continuum [98], we see humanoid robots in one end and virtual agents in the other end of the continuum, based on the character and scope of their embodiment in the physical and virtual worlds, respectively (Fig. 5). One main advantage of robots over embodied conversational agents (ECA) is their physicality that evokes a higher sense of presence, wider sensorimotor repertoire, and their ability to freely move to explore and interact with the physical world. On the other hand, ECA can demonstrate a strong sense of anthropomorphism due to their highly expressive and sociable interfaces [99] while amounting much less than a robotic interface. It is therefore no surprise to see mixed designs that meet at the middle of these two. However, researchers have also embodied their agents as intelligent ubiquitous and ambient realistic environments (left end of Fig. 5), augmented with networks of sensors, actuators and human–computer interfaces that provide an omnipresent controlling capability in small (e.g., smart homes [100]) to large and wide (e.g., smart cities [101]) scales. Thus, a research agenda is to study the bandwidth of communication's impact on the human perception of a synthetic sense of control and causal action [43].

The final and most important aspect of our conceptual model is the mediative role of Fp_H: as long as Fp_H is strong, the attitude toward AI will remain positive even if Fp'_{AI} is strong. Thus, it is crucial that the AI maintains knowledge via its Sp_{AI} of the changes in Fp_H (this is a significant part of, and which we elucidate later, in our third argument). The particular concern that our model addresses in this aspect is the creation of AI-enabled systems that are socially acceptable, and one way to study this is to focus on people's attitudes toward AI. We take for example investigations in human–robot interactions (HRI). Zafari and Koeszegi [103] discussed in length (with notable references) how people's acceptance of robots hinges on the technological, mental and societal aspects of robots interacting with humans in everyday life. According to them, there is higher probability for robots to be utilized when they seemed to have less agency, when people were more anxious toward a more adaptive over a less adaptive robot, and what foster user acceptance is not just the physical embodiment but also how the robot's appearance conforms to how it behaves. However, engagement in joint attention with robots is also more likely when robots are perceived as system with intentionality, and were attributed greater cooperativeness when they act according to team goals even to the point of users ignoring instances of their disobedience. Further in joint interactions with a robot, it has been shown that diffusion of responsibility occurs [85,104,105]. Diffusion of responsibility is a socio-psychological phenomenon whereby an individual is less likely to take responsibility for action or inaction when in the presence of other people. One fails to act by assuming that since other witnesses nearby are not acting, action is uncalled for and unsuitable. This lack of action and sense responsibility characterizes the lack of SoA [5,9,106]. More explicitly, findings show that our sense of responsibility can be diffused when others are around, due to the debilitation of the neural action-outcome link, which thus results in decreased SoA [104,105]. Based therefore on the findings of [85,92–94,103–105], among others, it can be said that SoA is reduced in HRI.

Zafari and Koeszegi [103], however, presented further a very interesting hypothesis: with the perception of the robot's agency increasing with the degree of its autonomy [107], relative to individuals exposed to non-agentic robots, people who encounter robots which they believe have more control than they should will report greater negative attitudes toward these robots. They cited several literature that support their proposition, such as increased robot autonomy correlates to negative emotions due to the users' perceived lack of control or when users see control to be with another entity, individuals who experience more control tend to be more comfortable cooperating with a robot that do not seem to undercut or threaten their values or significance to the task, and the reactance theory that people respond in a negative manner when their sense of control is threatened whatever the source may be. Zafari and Koeszegi's experiments in [103] proved their hypothesis to be correct.

Fig. 5. Bandwidth of communication. We situated examples of embodied agents in Milgram and Kishino's Continuum (L-to-R; sizes are not scaled: Intelligentenvironment, Sophia, Asimo, Nao, Pepper, Jibo, BellaBot, FURo-D, Greta [102] and Max [?]). We posit the perception of Fp'_{AI} by Sp_H is influenced by the bandwidth of communication afforded by the agent's embodiment in which this bandwidth increases with more natural kinds of abilities, including but not limited to perception, locomotion, object manipulation, speech, gesture, facial expression and display of social and affective states [95].

From the above, we surmise that attitudes toward AI can be expected negative when Fp_H is low. The conditions that diminish human sense of control and causal influence may result to decreased motivation and cognitive capacity, and consequently, negative reception of the AI. The concern therefore in the design of AI-enabled systems is how to keep Fp_H strong to continuously elicit positive attitudes toward AI. The goal therefore is for AI to adapt to dynamic changes in Fp_H in order to sustain it when strong and improve it when diminished, which brings us to the essence of our third and last argument.

To make our conceptual model conceivable, we look at how the concern on the human experience of control has been front and center in the acceptability of (attitude toward) automated and AI-enabled manufactured products. Before we dive into this, we can first trace earlier investigations on how an increased Fp_H positively affected consumer satisfaction and product acceptance in the service and production industries (see [108–114]). The tagline "Let Hertz put you in the driver's seat" was popularized in the 1950s by the car rental corporation Hertz, which catered to consumers' innate need for control. Convincing customers that they may have it their way (Sp_H) became a key component of competitive warfare in some retail industries (e.g., fast food restaurants). Providing customers with the opportunity to choose (Sp_H) and communicating this through ads on billboards or TV, foot sales agents, flyers, and the like, which we can say are these industries' bandwidth of communication (BoC), increased customers' Fp_H. In other words, when customers viewed themselves as in control, they would often take chances, and even risks, on services and products.

Came automation in consumerism. Consumer products ranging from vehicles to small and big appliances have been automated for the benefit of the user. The consumer operates (hence, the consumer becomes the operator) the appliances by simply pressing a series of buttons in the proper sequence and the machine then operates by itself (BoC), or by turning the steering wheel of an automobile. Decrease in perceived control (diminished Fp_H) when engaging with highly self-operating systems (BoC) can imperil (Sp_H) the approval of these systems' decisions by the human operators as we discussed above.

Finally, comes our conceptual model for the current state of the art: the unceasing advancement of automated systems, and particularly, the use of AI, which is progressing at a high rate and infiltrating many facets of modern life (BoC). As a result, there has been a surge of interest toward user acceptance of AI technology. The acceptability of AI is likely to be greatly influenced by people's overall opinions regarding it. For instance, in the majority of interactions between humans and machines, most of the system's automated decision processes and operations are black-box, i.e., unknown, inaccessible, or not explainable to the operator [115], and this observed lack of transparency (Sp_H) is a key factor that makes automation particularly hazardous to the operator's Fp_H. Lack of transparency makes it difficult for the operator to anticipate the transition of outcome events. As we discussed in Section 2, predictive mechanisms are known to be vital in the development of Fp_H. By disrupting the predictability of action-outcome events, the inherent lack of transparency of technological systems will therefore influence the Fp_H of human operators. Vantrepotte et al. [33] showed that the explicability effect, i.e., giving added information that could reveal more of the system's decisions and operations (BoC), was associated with stronger Fp_H and greater confidence in the decisions of the system (positive attitude toward the system) because the human operator understood more about the system (Sp_H). When it comes to AI, the black-box nature of deep neural network-based models is well-known, which has brought about the need for explainable AI [116,117]. What our conceptual model posits is that if the AI's capabilities (Fp'_{AI}) are made explainable (explicability effect) to the human (Sp_H), this would lead to increased Fp_H and create the positive attitude toward the AI. What the literature has shown is that the explanability of AI has led to trustworthy AI [118,119], but we posit that this is because the explicability effect in AI increased the Fp_H that led to the AI being trusted.

5. The third argument

Research on $Fp_H \rightarrow Sp_{AI} \rightarrow Fp_{AI} \rightarrow Action$, i.e., an AI that senses, makes sense of and responds sensibly to dynamic changes in human SoA, albeit compelling, is practically non-existent, and therefore needs to be realized.

Finally, we argue that what is gravely lacking, and practically non-existent, is research on how an artificial system, let alone one that is AI-enabled, can be SoA-cognizant to consequently support SoA. Per the two-person SoA perspective (Fig. 1(b)), this means the AI is able to *sense* Fp_H, *make sense* of Fp_H through its computational representation in Sp_{AI}, and *adapt sensibly* through its Fp'_{AI} (not necessarily an actual Fp_{AI}) for action generation and control to achieve causal influence in the world in order to sustain a strong, or improve a diminished, Fp_H. This is a non-trivial task, and we demonstrate in this section the extent of our argument by showing how few the exceptions that attempt to pave the way to realizing an SoA-cognizant and responsive AI.

Before we proceed further, it is worth understanding first that the interaction effect between Fp_H and Fp'_{AI} is not mutual for two reasons. As we detailed in Section 4, under the second argument, the effect of Fp'_{AI} on Fp_H is generally detrimental. The different research groups of Ciardo, Beyer, Zafari, Grynszpan, and Roselli, among others, have demonstrated that SoA is reduced in HRI, and even automation alone that is void of AI can reduce SoA as evidenced by the separate research findings of, inter alia, Berberian and Wen. Here is where the non-mutuality emerges: while the effect of Fp'_{AI} on Fp_H is generally detrimental, the AI as a machine, on the other hand, cannot be affected by Fp_H. Traditional AI is programmed to put what the human values (e.g., goals or objectives during tasks), normally represented as or part of an objective function, as secondary or even irrelevant and would perform to optimize only its own objective function. The second aspect of this non-mutuality is that the AI, by nature and with its computing power, can regulate its control more than the human. Indeed, the human can regulate interoceptively its SoA, but may generally find it difficult in the presence of automation or AI that can have full-control, as we have explained. The general essence of the third argument is that the AI should do something (or perhaps do nothing) to regulate its Fp'_{AI} when it senses the diminishing Fp_H, and focus its actions on stabilizing or improving the Fp_H (e.g., getting the human to trust its Fp'_{AI} by explaining its processes or returning the control back to the human).

5.1. Arguing for an Sp_{AI} : Modeling Fp_H in complex, natural settings

We argue that while there are only a handful of research works that have observed the dynamic changes in SoA, none yet has attempted to automatically model the dynamic behavior of human SoA in real-time in a natural, let alone, complex setting. This is the challenge imposed on *sensing* SoA: to detect and estimate SoA level changes through some representation of Fp_H in Sp_{AI}. To our knowledge, there has yet to be a computational model of Sp_{AI}.

To represent Fp_H in Sp_{AI}, first, there needs to be a viable mechanism to measure Fp_H naturally in real-time. Investigating SoA has proved difficult to assess since we are most of the time not aware of, let alone intentionally inquiring about, our SoA each time we act [1]. Traditional methods in psychological experiments measure SoA either explicitly or implicitly to detect and estimate SoA levels (see reviews in [6,120]). Explicit measures involve the subjective estimation and reporting of one's own SoA over a specific task or event. Rating scales are commonly used for participants to report their agreement, e.g., between 1 (definitely not) and 10 (absolutely), to inquiries of the form "Did you/your action cause that thing to happen?" On the other hand, implicit measures quantify the self-reported perceived differences between actions that were self-generated versus externally-generated without directly examining their SoA, notably, the intentional binding and sensory attenuation effects (see reviews in [71,121]). Intentional binding (IB), for instance, has been reported as a robust and dependable implicit measure of SoA, providing compelling analyses for many investigations on the temporal perception of action-outcome effects as it relates to the nature of SoA. The IB refers to how the perceived action (e.g., pressing a button) and its external sensory outcome (e.g., a tone sounding after the button-press) are attracted together in time, i.e., the action seems to have happened later than it actually did, while the sensory effect seems to have happened earlier in time, hence the binding effect. This phenomenon occurs only when actions are voluntary, thus, when SoA is experienced (conversely, the time between the action and outcome is longer when no SoA is involved). The IB, as well as sensory attenuation, however, have limitations: self-reporting by the participants involves continual interruption of their actions, and the tasks in which SoA had been examined were simple and confined in very controlled experimental settings. These make these methods less viable when dynamic changes in SoA need to be assessed unobtrusively in real-time, natural setting.

Wen and colleagues [30] offer three potential approaches to overcoming the limitations of the explicit and implicit measures. First is the physiological suppression of alpha-mu rhythm when detected by a brain-computer interface (BCI) that is linked to SoA during movements. The basic limitation of BCI, however, is that it is intrusive and the signals received from the brain are highly prone to interference, which most of the time requires subjects to do minimal movements. Another is the attention measured from eye gaze behavior, which indicates people are unconsciously more attentive to objects that would give them more control. The common constraint for the eye gaze tracker is that it requires the camera to be always properly aligned with eyesight. However, it is difficult to control eye position accurately all the time that results to the eye tracker providing unstable output when it does not get appropriate images of the eye in sequential frames. The third is a probing mechanism, i.e., generating a stimulus that warps one's prediction of the outcome, which would then elicit a response to correct the prediction — this recovery behavior can only arise when there is still SoA. This last method, however, does not detect nor estimate when, but only probes whether SoA is disrupted. Moreover, forcing a perceptual prediction error may also introduce interruptions to the natural intentional action flow.

Legaspi and colleagues [122,123], however, suggest a data-driven approach: SoA levels can be inferred from sensor data in the same manner as human affective states. It has been suggested that our affect (i.e., emotion or mood) interact with our SoA in a myriad of ways, with different possible mechanisms mediating the interaction of affect and awareness of action [54,124]). Here is the plausible bridge, Legaspi et al. suggest that SoA changes can be modeled in the same way the field of Affective Computing has successfully modeled human emotion, i.e., through bodily (physiological, facial, vocal, postural and gestural, among others) signals with the subjective self-assessment of their experience of agency as ground truth. Such bodily signals have been shown to accurately depict changes in affective states [125]. Human SoA, Fp_H, can therefore be modeled by mining the streams of multimodal behavioral signals captured by sensors that are built into smart phones, watches, or homes as people go about their daily lives and labeling these recorded signals with the self-reported SoA levels (e.g., in [123] are smartphone versions of assessing general explicit SoA [120] and implicit SoA via intentional binding [65]). Once an Fp_H model is generalized, it can then be used in normal living settings in which the SoA can be predicted by the model in real-time without any more asking the human to self-report her SoA while engaged in daily tasks. This Fp_H modeling mechanism therefore constitutes the AI's Sp_{AI} in our computational framework (see Fig. 6(top)).

Computationally, we can model the Fp_H with an estimation function that outputs the estimated Fp_H state value, that is,

$$\Psi : \delta_{Ph} \times \delta_{Po} \times \delta_G \times \delta_V \rightarrow \hat{Fp}_H \in \mathbb{R}^+, \tag{2}$$

where the δ features represent the inferred physiological, postural, gestural and vocal prosodic behavioral phenotypes, respectively, of the human from mining the streams of multimodal behavioral signals captured by wearable and ambient sensors. The δ values shall be fed into the recognition function Ψ to train the Fp_H model with the actual Fp_H value as ground truth to compare with the estimated $\hat{Fp}_H$. The model shall be trained to minimize the difference across all estimated and actual SoA values in the training dataset, i.e., $|\hat{Fp}_H - Fp_H|$. Metrics that are based on this difference can be employed. For instance, the root mean squared error, which is one of the most widely used measures to judge a model's performance, whether it be during training, cross-validation, or monitoring after deployment, can be computed as $RMSE = (\frac{\sum_{i=1}^{n}(\hat{Fp}_{H_i} - Fp_{H_i})^2}{n})^{1/2}$, where n is the number of training data samples. The estimation function Ψ can be trained until it has generalized such that $\hat{Fp}_H$, based on the chosen metric(s), has met the target accuracy and the model is therefore deemed suitable for deployment. Once deployed, the Fp_H model will be used by the SoA-cognizant AI to get informed as to whether Fp_H is diminishing, and will then perform the suitable Fp_H-centered interventions, which we explain next.

5.2. Arguing for Fp'_{AI}: Making sense of and adapting sensibly to Fp_H

Like Sp_{AI}, there has yet to be a computational model to make sense of and respond adaptively to human SoA, i.e., a model of Fp'_{AI} has yet to exist. However, the conceptualizations of Legaspi et al. in [122] (also adapted in [123]), as shown in Fig. 6(bottom), is a recent attempt toward realizing this in the context of their *S*oA-aware *P*ersuasive *Ar*tificial Intelligence (SPEAR). SPEAR is formulated as interacting with its human counterpart in a world with causal structures, and together they devise plans to optimally go about in the world while sharing the same incentive mechanism in which SoA is the most important component. The sets of states, actions that can be taken from those states, probabilities of transitions between states, likelihoods of possible actions, and probabilities of observations form SPEAR's belief about the world that it can partially observe and deduces from it the plausibly existing causal patterns. SPEAR's tool would be a structural causal model that allows testing for three aspects of causal inferences, namely, associative, interventional and counterfactual reasoning [126,127]. As a result, SPEAR can discern the action causal outcomes that originated from the human or itself.

Legaspi et al. further posited a scenario where SPEAR devises a new path to the desired destination, or to an unanticipated alternative

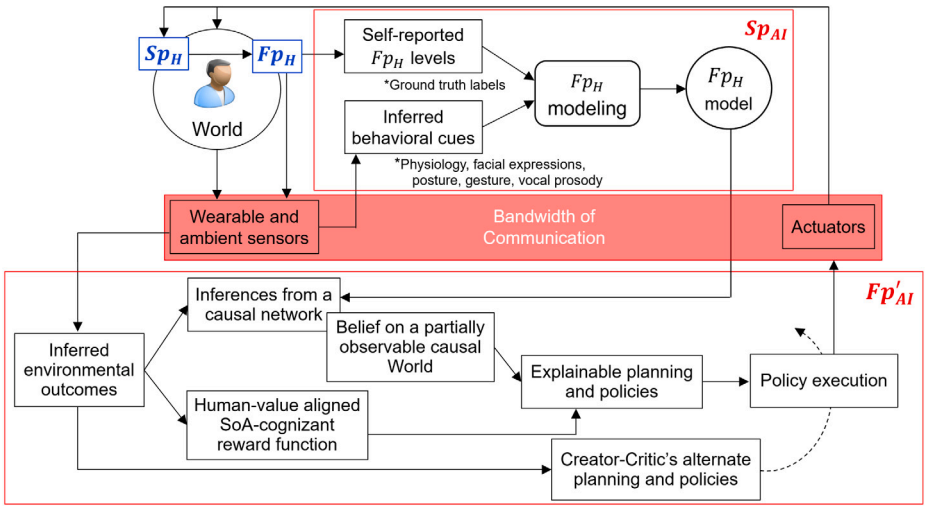


Fig. 6. Augmented computational framework. **(Top)** The computational model for Sp_{AI} as elucidated in Section 5.1. **(Bottom)** We posit that the framework in [122] of an AI-enabled persuasive system that is SoA-cognizant can be employed to model Fp'_{AI}.

but more secured destination, due to heavy traffic on the road or a roadblock caused by a typhoon or disaster. In other words, it would be advantageous for SPEAR to produce fresh or innovative agentic chances in unanticipated, unforeseeable, or uncommon circumstances. In contrast to what is ostensibly available in the current situation, SPEAR would derive a novel strategy that was initially unknown to both itself and the human. To achieve this, SPEAR would employ its Creator-Critic mechanism in which the Creator generates a counterfactual version of reality that can provide an alternative solution, e.g., a way to reaching a new target state assuming such state can be characterized by some exogenous source (e.g., citizen sensors), which can also be a feature of its bandwidth of communication. But as SPEAR learns a new strategy that derives previously unknown states, it uses the Critic module to validate this strategy. By adapting what it deems to be the more effective plan, SPEAR would offer new tactics that are more useful when utilizing shifting opportunities and assisting the human to perform optimally. Thus, SoA will increase as the AI provides fresh and innovative agentic opportunities to positively influence and guide human behavior.

We posit that the computational framework of SPEAR can be augmented with the components of our conceptual model. Given SPEAR's computational abilities, we suggest the bandwidth of communication can therefore refer to more than the embodiment, but also to the capabilities afforded in SPEAR for it to adapt sensibly to Fp_H so that the human via its Sp_H can perceive SPEAR as cognizant of human SoA (mid-section of Fig. 6) and therefore reinforces for strong Fp_H and positive relationship with SPEAR, following our conceptual model (Fig. 4). Armed with perceptual (e.g., wearable and ambient sensors) and actuating (e.g., robot or ECA interaction capabilities, smart home automatically changing temperature and light settings) devices and algorithms, and by farming through knowledge made available by the Internet-of-Things, the AI can inform the human with increased certainty of the situation that envelopes them both. As research works have shown, precise perception and integration of contextual cues increase SoA (Section 2). We consider SPEAR's perceptual and actuating devices and algorithms, and IoT knowledge it can efficiently acquire, can consist its bandwidth of communication.

The final thing we want to discuss is that Legaspi et al. theorize the various components of their framework can be mapped computationally to the sequential decision-making processes that involve the planning, selection, execution and policy construction of the actions of an explainable AI (XAI) [128,129]. They outlined the formalism by which the AI makes its plans and inferences transparent to the user and enables the user to show, challenge, and ask questions about them. The AI can become more effective in persuading the human (e.g., to change behavior) if it is able to make transparent and explain the rationale behind its behavior, may improve human and AI relationship as the AI is perceived to be trustworthy, and lastly, as it conveys to the user its comprehension of human behavior, it might be seen as being empathetic. The human's SoA and the positive attitude toward SPEAR are suggested to increase as a result of these procedures that enable the human to co-produce with SPEAR the intentional actions to achieve the desired outcomes.

5.3. Empirical validations on estimating the influences of and on SoA

In [122], Legaspi et al. illustrated how their computational framework, Fig. 6(bottom), could be implemented in a real-world scenario employing as illustrative use case a study [130] that looked at how emotional persuasion affected drivers' detour behavior. We reckon that actual field experiments should be conducted to prove every aspect of their computational framework, and empirically validating it is indeed a noteworthy future task for anyone who desires to pursue it. Nonetheless, we show in the succeeding sections how quantitative methods and metrics are applied to demonstrate that SoA can be captured, recorded, estimated and modeled, thereby demonstrating Sp_{AI} modeling, Fig. 6(top). Furthermore, if sensing, making sense of and acting sensibly to influence SoA is indeed at the core, it is essential for an SoA-cognizant system to accurately estimate Fp_H so as to be influenced by it and to positively influence it. We discuss below the results and analyses of two experiments that demonstrate how the influences of and on SoA can be estimated to support human SoA.

5.3.1. Recording and analyzing Fp_H behavior in a natural, complex setting

Given the above, the first question is whether changes in SoA can be captured and recorded in a natural, real-world scenario, i.e., outside the auspices of a controlled lab environment wherein the participants can perform the experiments naturally with least constraints as possible. This is a non-trivial task, and none yet has attempted to automatically model the dynamic behavior of human SoA in real-time in a natural, let alone, complex setting. With the exception perhaps of the work of Legaspi and colleagues, which we discuss below.

Increased agency has been observed to propel one's self to succeed in controlling its own behavior while in pursuit of a defined goal [131], and that the experience of self-agency depends on maintaining a mental representation of that goal [132]. To investigate SoA this way, Legaspi et al. [123] defined healthy eating behavior (essential to physical and mental well-being [133,134]) as the goal, then formulated and carried

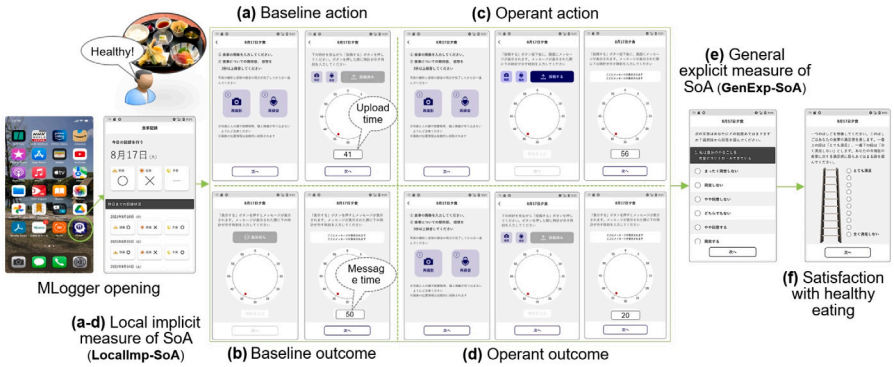


Fig. 7. MLogger shown in [123]: The mobile app used by Legaspi et al. to quantify SoA in a multidimensional manner and in a natural setting on a daily basis (3x/day). We refer the reader to their paper [123] for the complete description of all the features of MLogger.

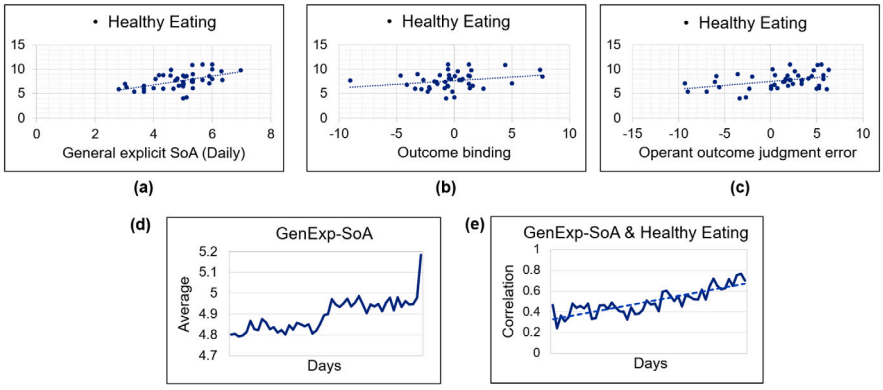


Fig. 8. Results and analyses in [123]. The top plots show the daily components of the general explicit (a) and local implicit (b and c) SoA that have significant direct effects on healthy eating behavior. The bottom plots show the subjects' aggregated daily perceptions of their general explicit SoA (d) and together with their healthy eating behavior (e).

out a multidimensional investigation of SoA to determine how a general explicit SoA (GenExp-SoA for brevity) and local implicit SoA (LocalImp-SoA) could influence healthy eating over time, and whether a relation exists between GenExp-SoA and LocalImp-SoA. To aid in their research, they constructed and deployed their own mobile application, called MLogger ("M" for meal), shown in Fig. 7. While it can be seen in the figure that there are many dimensions in Legaspi et al.'s multifaceted investigation, note that what is pertinent to this section is to highlight how the influence of SoA can be recorded and analyzed. We refer the reader to the details of their experiments, data, results, and analyses from using the MLogger in [123].

The MLogger begins with an improvised intentional binding set-up, one that is specifically designed for a smartphone, to measure the LocalImp-SoA (Fig. 7(a) to 7(d)) on healthy eating (Fig. 7(f)). They also measured the participants' GenExp-SoA using a subset of the general SoA Scale [120] in two occasions, prior to the participants engaging in the actual experiments (i.e., pre-experiment) and then daily thereafter with the participants using MLogger (Fig. 7(e)). The experiments ran for 42 days with 43 participants, through which a total of 5215 data points had been collected.

Legaspi et al. analyzed all the data they collected from the pre-experiment surveys and daily use of MLogger. Using linear regression analyses, their results showed that while there were no direct effects of the pre-experiment GenExp-SoA, the daily GenExp-SoA had an effect on healthy eating behavior (Fig. 8(a)). Moreover, when the participants' daily LocalImp-SoA was analyzed, Legaspi et al. found an effect of the outcome binding components on healthy eating (Fig. 8(b) and (c), respectively). Analysis using ANOVA indeed showed participants' daily GenExp-SoA varied, indicative of the their subjective experiences of GenExp-SoA. Daily averages of their GenExp-SoA also varied (Fig. 8(d)), and the daily mean correlations of their GenExp-SoA and healthy eating increased together in the course of the experiments (Fig. 8(e)). All these suggest that the participants' SoA changed over time, and that the influence of SoA on the goal as it increased over time throughout goal pursuit can lead to achieving the goal. Lastly, despite the fact that prior research has only extensively examined local explicit and implicit SoAs and observed no association between them (references in [123]), Legaspi et al.'s multidimensional analysis actually found no significant correlation, nor any interaction effects, between GenExp-SoA and LocalImp-SoA. Contributing therefore to prior observations, they discovered GenExp-SoA and LocalImp-SoA are also not related.

The analyses above of Legaspi et al. suggest the viability of assessing SoA that changes over time, recognizing its behavior and how it impacts goal pursuit. When SoA is detected to have decreased, which could impede achieving the desired consequence, suitable intervention can be performed to alter how oneself perceives its agency. Legaspi et al. explained in [123] how their analyses is linked to user-adaptive AI interventions by adapting the pertinent components of their SoA-aware persuasive AI in [122] (also discussed in 5.2).

5.3.2. *Modeling the influence on* Fp_H

We conducted our own experiment, which employed a dataset that was collected from 54 subjects (male: 34 and female: 18; age: min=29, max=49, average=41.0; with normal hearing and vision) when they self-reported their SoA states as they listened to different musical pieces and their brainwave productions monitored by an EEG device. We conducted our data collection from Nov. 11 to Dec. 22 of 2022 with the approval of the Ethics Committee of our organization.[3] All subjects

[3] It was paramount that we performed ethically our experiments to intentionally safeguard the privacy and dignity of all our subjects. Subjects

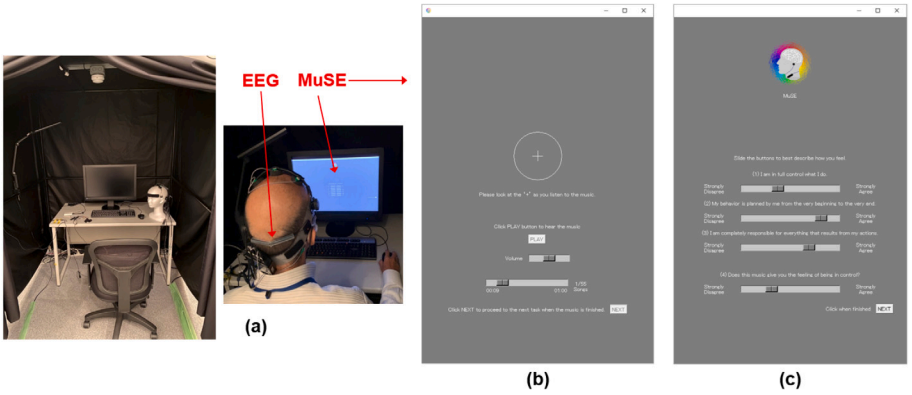


Fig. 9. Experiment with MuSE. While performing the experiment (a) the subject is seated in an enclosed cubicle, wore an EEG and used the MuSE data collection app. (b) Music started playing at the time the subject decided to, and (c) the subject reported his SoA by using a visual analog scale.

were fully informed ahead of time of the methods and contents of the experiment, and subjects gave their consent before the experiment was conducted on them. We asked the subjects to come to our experiment facility in which they were seated individually in an enclosed cubicle (Fig. 9(a)) so as not to be distracted by other exogenous factors and to focus only on the experiment.

To prepare the music playlist, we collected an initial set of 143 copyright-free music files (no lyrical contents), each with 60-s duration and categorized according to their musical flavor, i.e., motivational, suspenseful, horror, sad, energetic, or uplifting. We then internally assessed each piece and voted on which ones to include in the experiments. We removed those that seemed to have very similar flavor contents, as well as those that we found boring and could make the subjects less engaged (e.g., sleepy). At the end, 60 musical pieces comprised the playlist.

During data collection, subjects were asked to follow the instructions given by our own desktop application (hereafter, MuSE; Figs. 9(b) and (c)) that played the music stimuli and recorded the subjects' self-reported SoA ratings. We asked the subjects to fix their eyes on the cross that is in the middle of the screen (Fig. 9(b)) to prevent noise from being included in the EEG (Fig. 9(a)) data due to eye movements. MuSE randomly selected the musical pieces from the playlist and played them for the subjects. The number of music stimuli the subjects heard varied (min=17, max=51, average=35.6) depending on each subject's pace. Thus, the subjects did not necessarily hear the same set of music stimuli. After listening to each musical piece, using MuSE's visual analog scale, the subjects indicated the degree of their responses to four aspects of SoA ((Fig. 9(c)), which were "(1) I am in full control of what I do"., "(2) I am completely responsible for everything that results from my actions"., "(3) My behavior is planned by me from the very beginning to the very end"., which were taken from the general SoA Scale [120], and were actually the same items used by Legaspi et al. in [123] to measure the daily GenExp-SoA, and "(4) Does this music give you the feeling of being in control?" The subjects moved the slider of the visual analog scale between "Strongly Disagree" and "Strongly Agree". In the preprocessing stage, we converted the scale values to between 1.0 and 15.0 respectively.

Our theory is that the different music stimuli would affect differently the subjects' SoA. For each musical piece that a subject listened to, we averaged the ratings on the SoA-related questions (1), (2) and (3) above, and treated this average (hereafter, Music-SoA) as the subject's SoA that might had been affected by music. Legaspi and colleagues have shown in the past that music can elicit different affective states in listeners [135,136], and as we mentioned in Section 5.1, it has been argued that our emotion or mood interacts with our SoA in different ways under different mechanisms that mediate their interactions. Thus, we posit music can influence SoA as much as it influences human affect. To confirm this, we plotted the mean of each subject's Music-SoA (Fig. 10(a)) and observed from these that even though Music-SoA basically measures the same aspects each time (i.e, (1), (2) and (3)), Music-SoA may differ within subjects. We confirmed this even more when we plotted the mean of the responses to question (4) (Fig. 10(c)). These results suggest that what affected the change in SoA is not the SoA-related life situations the subjects were into at the time of the experiments, rather, the musical pieces elicited the changes in their SoA. Further analysis using ANOVA also revealed that SoA varied across participants ($F[53, 2081] = 110.848, F_{Crit} = 1.346, p < 0.001$ for Music-SoA, and $F[53, 2081] = 26.400, F_{Crit} = 1.346, p < 0.001$ for (4)), indicative of their subjective feelings of self-agency after listening to the music stimuli. We also plotted the mean Music-SoA and responses to (4) across all participants for each musical piece (Fig. 10(b) and (d), respectively) and observed that the music stimuli had varying effects on the subjects' SoA based on their responses to (4) ($F[59, 2075] = 4.022, F_{Crit} = 1.328, p < 0.001$), albeit not in Music-SoA ($F[59, 2075] = 0.879, F_{Crit} = 1.328, p = 0.731$). All these suggest that music had individualized, varying influence on the subjects' experience of self-agency.

We then created two kinds of SoA-predictive models, one that predicts SoA from music features and the other predicts SoA from EEG features. For the fist model, we segmented each raw 60-s music signal in WAV format using an overlapping sliding window, i.e., a 3.0-s window with 1.5-s overlap size slid through the raw signal. With a sampling rate of 22,050 Hz, every music signal contained 1.323 million data samples, and each music-segment window contained 66,150 samples. We then extracted from each window 487 features using librosa,[4] commonly used by the music information retrieval community. For the second model, we segmented each 60-s EEG signal in the same manner as that of music for them to be synchronized, which resulted to 30,000 and 1500 data samples for the entire EEG signal and EEG-segment window, respectively, given a 500 Hz sampling rate. Using now eeglib[5] [137], we extracted 1540 EEG features from each EEG window-segment. Finally, we labeled with each subject's Music-SoA as ground truth the synchronized music and EEG feature vectors. We enumerated in Table 3 the set of extracted music and EEG features, as well as the number of dimensions for each feature.

We trained both prediction models and validated their performance. We used for training the music and EEG feature vectors, together with

participated on their own volition, and any personally identifiable information (PII) was removed before we acquired the data.

[4] librosa is sound processing library for python; https://librosa.org/doc/latest/index.html.

[5] The eeglib module is a library with tools and functions for EEG signal analysis; https://eeglib.readthedocs.io/en/latest/.

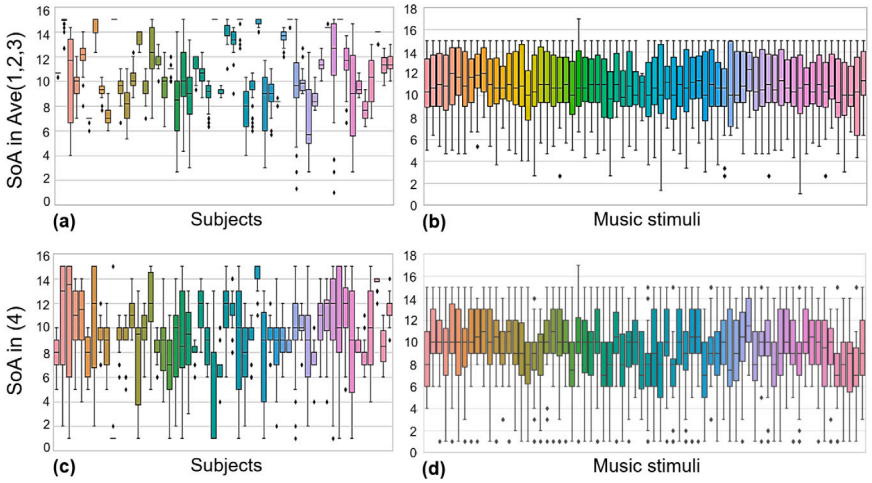


Fig. 10. Influence of music on SoA. The graphs show that the influence of music on the subjects' SoA varied, as shown by each subject's self-reported SoA levels in response to questions (1) to (3) in (a) and (b), and to question (4) in (c) and (d).

Table 1
SoA from music. Performance of the different regression models when predicting SoA from music features.

	LinearReg	XGBoost	RBF-SVM	CatBoost	Stacked LSTM
RMSE	2.992	2.832	2.830	2.818	**2.768**
MSLE	0.078	0.069	0.069	0.070	**0.067**
MAE	2.463	2.330	2.325	2.331	**2.273**
MAPE	26.566	25.345	25.425	25.941	**25.062**
LogHypCos	1.864	1.737	1.732	1.864	**1.687**

Table 2
SoA from brainwave signals. Performance of the different regression models when predicting SoA from EEG signals.

	LinearReg	XGBoost	RBF-SVM	CatBoost	Stacked LSTM
RMSE	2.679	**2.503**	2.621	2.685	2.616
MSLE	0.064	**0.056**	0.061	0.065	0.061
MAE	2.197	**2.001**	2.104	2.197	2.157
MAPE	26.409	**21.806**	23.272	24.503	23.860
LogHypCos	1.607	**1.435**	1.529	1.614	1.569

the corresponding Music-SoA labels, that were associated to 47 music stimuli, and the rest of the data associated to 13 music stimuli for testing. Since the Music-SoA ground truths are continuous values, we trained several predictive regression models, namely, linear regression, XGBoost, Radial Basis Function (RBF) SVM, CatBoost, and a stacked LSTM (with three LSTM layers trained for 800 epochs). Prior to training the models, we used interpolation to fill the missing feature values and then applied normalization. We evaluated the performance of the models using regression metrics, namely, mean squared error (RMSE), mean squared logarithmic error (MSLE), mean absolute error (MAE), mean absolute percentage error (MAPE), and log hyperbolic cosine (LogHypCos). Our results on predicting SoA from music features (Table 1) show that the more advanced stacked LSTM regression model outperformed the other ML methods. More importantly, this shows that SoA levels can be accurately predicted with relatively high accuracy, for instance, the accuracy of prediction is around 74.94% based on the MAPE for stacked LSTM. However, we can observe that when predicting the SoA from EEG features, XGBoost outperformed LSTM (Table 2). Even more, the accuracy of this model's predictions is higher across all metrics than when using the music features (cf. Table 1), for instance, with a prediction accuracy of about 78.19% based on MAPE.

What we have shown above is that the influence of the stimuli on Fp_H can be estimated with relatively high accuracy from both music and EEG features. Furthermore, predicting the influence of the stimuli on Fp_H using EEG features is more accurate given the physiological connections between SoA and the neuro-cognitive functions. [4,5]. This demonstrates our estimation function for Fp_H in Eq. (2) in which $\Psi : \delta_{EEG} \rightarrow \hat{Fp}_H \in \mathbb{R}^+$.

5.3.3. Tying the conceptual, theoretical and empirical knots

We now put the pieces together in light of our conceptual model. We have shown that Fp_H can be captured in a natural setting, e.g., using applications (MLogger and MuSE) through which users could self-report their SoA. We also demonstrated that the captured Fp_H, together with the characteristics of the stimuli (music), can be modeled in Sp_{AI} using conventional statistical analyses (per [123]) or machine learning methods (per our results). Moreover, through a wider bandwidth of communication in the AI, Sp_{AI} can also model the effect of the stimuli on Fp_H by modeling the SoA-related physiology (brainwave) signals that are activated when the stimulus is applied. We should repeat here, however, that the limitation of brain interfaces is that the signals received from the brain are highly prone to interference that could require subjects to do minimal movements. If SoA can be estimated from physiology signals while users are engaged in daily life routines (e.g., using MLogger to collect physiology-related signals, such as heart rate variability, skin conductance and temperature), it is no longer necessary for the users to self-report their SoA, which could distract them from their usual activities. Instead, the levels of SoA can be automatically predicted from the streaming physiology signals.

What we posit in light of the above is that we can envision target goals in which there is a need for the user to maintain a strong sense of self-control when pursuing a goal, such as smoking cessation, healthy diet, overcoming phone addiction, or good study and working conditions. If the user's Fp_H can be captured and predicted in real-time by an AI through its Sp_{AI}, and if it evaluates Fp_H to be decreasing, then the AI can use an interventional stimulus, such as music, as component of its Fp'_{AI} to help increase or improve the user's self-control. In other words, the AI will play its predicted SoA-increasing music when it senses the user's low Fp_H.

The essence of the third argument is that the AI should intervene via its Fp'_{AI} when it senses the diminishing Fp_H, and focus its actions on stabilizing, regaining or improving the Fp_H (e.g., returning the control back to the human per Section 1 or getting the human to trust its Fp'_{AI} by explaining its processes per Section 4). The attitude of the user when interacting with the AI is paramount. Ultimately, as it is the core of this paper, the AI should be accurate when demonstrating its Fp_H-aware capabilities, and these capabilities should be assessed based on the observable proxy behavioral variables that quantify the manifested effects of the latent Fp_H variable (cf. Eq. (2)).

The approach we took above in implementing the prediction models can be viewed as offline learning, in which a global cost function is reduced while the model weights and parameters are updated during training. The model is trained until it satisfies the set criteria for deployment or use case that it has been designed for. The models can be evaluated based on several criteria as we have shown, but may also be in terms of the algorithmic time and space complexity, scalability of the models, and their use in production. The caveat with offline learning is that the models must have seen relatively sufficient data (see also few-shot learning) to represent the real word. But the real world is noisy, uncertain, and there may not be any low-dimensional model that can be fit to its high-dimensionality [138,139]. We envision an AI-enabled system that is able to adapt to changing user feedback in dynamic, natural, and complex daily environmental settings. Humans commonly make subconscious predictions about outcomes in the physical world and are surprised by the unexpected. In such a case, the model may well learn online, as it is continuously exposed to fresh data and is able to continuously improve through online learning.

The Sp_{AI} needs to become robust and general enough that when the system is deployed it does not need to ask the human each time regarding the state of her Fp_H. However, the Sp_{AI} must also account for changes in and from the user and environment. What can happen is that the model is first trained offline using data that is expected to be present in the real world. However, to account for the changing user and environment, a possible solution is active learning [140] whereby the model is able to query a user operator online in order to resolve any ambiguity during the learning process.

Finally, when evaluating the performance and impact of an SoA-cognizant AI as a whole and as a live system, online evaluation metrics can be used to assess its quality and effectiveness when interacting with real users and environments. Different online evaluation metrics can be used, which can be as general as user satisfaction metrics through self-reported ratings. A/B testing is a commonly used method where users are split into two or more groups, with each group exposed to a different version of the AI-enabled system (e.g., SoA vs. not SoA-aware, with vs. without online learning). Multi-armed bandits [141] is another method where users can be dynamically allocated to different versions of the system based on observed online evaluation metrics and a learning algorithm. Both of these methods can help test the causal effect of the SoA-predictive models in terms of online performance and impact.

The above tells us that an AI recognizing changes in human SoA and adapting its responses to improve the SoA is a compelling research direction. If we want to design the kind of technology we are suggesting, we would need to explore a lot of an unchartered terrain. First, the AI must have access to sensors, actuators, and learning algorithms in order to autonomously see the world, create adaptive models, and act in it. All these constituting the bandwidth of communication. The AI should be viewed as a collaborative partner whose main goal is to benefit humans by sharing their aspirations, principles, and values. The AI should be created in such a way that it facilitates inputs that enable humans to query or contest its decision-making processes. Finally, and most significantly, the AI should be aware of how human SoA evolves through time in response to the complex settings of a natural environment. This would necessitate sensitively asking the human to estimate this intrinsic behavior. But for this to occur, the AI must preserve a relationship with the human that is based on trust by being transparent on how it makes decisions and puts premium on human SoA.

6. Conclusion

We propound our work in this paper as taking an argumentative position, by attempting to capture in three arguments the germane body of knowledge on how human sense of agency is and should be affected by an AI that is perceived to possess an innate sense of control. We also hope to provide a solid starting point, a first clear outline in the literature, for researchers who are interested in beginning their research on sense of agency in human–AI interactions and to be a source of information for academicians and practitioners looking for state-of-the-art evidence to guide their work on this topic, which we achieved by analyzing and synthesizing pertinent works in the neuro, cognitive, and behavioral sciences, as well as in theoretical and computational AI.

The need to have control over the technologies we use is quite a powerful feeling in us as humans [26]. However, if an AI is believed to have a subjective perception of its own agency, this could be viewed as a disincentive to our own SoA since we may feel as though the AI have us under its control. As a result, the conundrum of joint human and synthetic agency would be included in the seemingly endless discussion of how humans and AI may interact wrongly, with the key idea being that an AI with a high level of agency can result in a dystopian future. But this should not be the case in the future since, as we argued and elucidated extensively in this paper, the design of AI-enabled systems can readily draw knowledge from the cognitive, neural and behavioral sciences on how they can be knowledgeable of and sensitive to human SoA by adaptively responding to sustain or improve it, rather than diminish or erode.

The goal of this manuscript is to argue our position on how SoA in human–AI interactions should be viewed and pursued. By farming through the SoA literature in the neuro, cognitive, and behavioral sciences, as well as in the technical fields of automation, HCI, HRI, HMI, and AI, we incorporated supportive evidence in the literature, rooted in facts, that provide a solid foundation for our arguments.

Table 3
Predictive modeling features used in Section 5.3.2.

Music features	Dim.	EEG features	Dim.
Zero-crossing rate	1	Power spectral density	1492
Root mean square	1	θ frequency band	4
Constant Q-transform (CQT)	84	δ frequency band	4
Chroma short-time Fourier transform	12	α frequency band	4
Chroma CQT	12	β frequency band	4
Chroma energy normalized	12	Hjorth activity	4
Mel-scaled spectrogram	128	Hjorth mobility	4
Mel-frequency cepstral coefficients	20	Hjorth complexity	4
Spectral centroid	1	Detrended fluctuation analysis	4
Spectral bandwidth	1	Sample entropy	4
Spectral contrast	7	Lempel–Ziv complexity	4
Spectral flatness	1	Petrosian fractal dimension	4
Spectral roll-off	1	Higuchi fractal dimension	4
Poly features	2		
Tonnetz	6		
Tempogram	30		
Harmonic CQT	84		
Percussive CQT	84		

CRediT authorship contribution statement

Roberto Legaspi: Investigation, Conceptualization, Methodology, Experimentation and validation, Writing – review & editing. **Wenzhen Xu:** Conceptualization, Review. **Tatsuya Konishi:** Conceptualization. **Shinya Wada:** Conceptualization. **Nao Kobayashi:** Experimentation and validation. **Yasushi Naruse:** Experimentation and validation, Review. **Yuichi Ishikawa:** Conceptualization, Experimentation and validation, Review, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was partially supported by the Innovative Science and Technology Initiative for Security, Grant Number JPJ004596, ATLA, Japan.

Appendix A. Acronyms

SoA	Sense of agency
AI	Artificial intelligence
Fp_H	Human's innate SoA
Fp_{AI}	AI's innate SoA
Fp'_{AI}	Humanly-perceived AI's innate SoA
Sp_H	Human's perception of the AI's SoA
Sp_{AI}	AI's perception of the human's SoA
HCI	Human–computer interaction
HMI	Human-machine interaction
HRI	Human–robot interaction
CCE	Confidence in causal estimate
IB	Intentional binding
ECA	Embodied conversational agent
BoC	Bandwidth of communication
SPEAR	SoA-aware Persuasive Artificial Intelligence

References

[1] S. Gallagher, Philosophical conceptions of the self: implications for cognitive science, Trends Cogn. Sci. 4 (2000) 14–21, http://dx.doi.org/10.1016/s1364-6613(99)01417-5, URL https://pubmed.ncbi.nlm.nih.gov/10637618/.

[2] S. Gallagher, Multiple aspects in the sense of agency, New Ideas Psychol. 30 (2012) 15–31, http://dx.doi.org/10.1016/j.newideapsych.2010.03.003, URL https://www.sciencedirect.com/science/article/pii/S0732118X10000218.

[3] N. David, S. Obhi, J.W. Moore, Editorial: Sense of agency: examining awareness of the acting self, Front. Hum. Neurosci. 9 (2015) 310, http://dx.doi.org/10.3389/fnhum.2015.00310, URL https://www.frontiersin.org/article/10.3389/fnhum.2015.00310.

[4] J.W. Moore, What is the sense of agency and why does it matter? Front. Psychol. 7 (2016) 1272, http://dx.doi.org/10.3389/fpsyg.2016.01272, URL https://www.frontiersin.org/article/10.3389/fpsyg.2016.01272.

[5] P. Haggard, Sense of agency in the human brain, Nat. Rev. Neurosci. 18 (2017) 196–207, http://dx.doi.org/10.1038/nrn.2017.14, URL https://www.nature.com/articles/nrn.2017.14.

[6] J.A. Dewey, G. Knoblich, Do implicit and explicit measures of the sense of agency measure the same thing? PLoS ONE 9 (2014) http://dx.doi.org/10.1371/journal.pone.0110118, URL https://doi.org/10.1371/journal.pone.0110118.

[7] M. Balconi, The sense of agency in psychology and neuropsychology, in: M. Balconi (Ed.), Neuropsychology of the Sense of Agency: From Consciousness to Action, Springer Milan, Milano, 2010, pp. 3–22, http://dx.doi.org/10.1007/978-88-470-1587-6_1, URL https://link.springer.com/chapter/10.1007/978-88-470-1587-6_1.

[8] Z. Barlas, S. Obhi, Freedom, choice, and the sense of agency, Front. Hum. Neurosci. 7 (2013) 514, http://dx.doi.org/10.3389/fnhum.2013.00514, URL https://www.frontiersin.org/article/10.3389/fnhum.2013.00514.

[9] E.A. Caspar, J.F. Christensen, A. Cleeremans, P. Haggard, Coercion changes the sense of agency in the human brain, Curr. Biol. 26 (2016) 585–592, http://dx.doi.org/10.1016/j.cub.2015.12.067, URL https://www.sciencedirect.com/science/article/pii/S096098221600052X.

[10] V. Chambon, N. Sidarus, P. Haggard, From action intentions to action effects: how does the sense of agency come about? Front. Hum. Neurosci. 8 (2014) 320, http://dx.doi.org/10.3389/fnhum.2014.00320, URL https://www.frontiersin.org/article/10.3389/fnhum.2014.00320.

[11] M. Synofzik, M. Voss, Disturbances of the sense of agency in schizophrenia, in: Neuropsychology of the Sense of Agency: From Consciousness to Action, Springer Milan, Milano, 2010, pp. 145–155, http://dx.doi.org/10.1007/978-88-470-1587-6_8, URL https://doi.org/10.1007/978-88-470-1587-6_8.

[12] M. Synofzik, M. Voss, B. Michela, Disturbances of the sense of agency in schizophrenia, in: Neuropsychology of the Sense of Agency: From Consciousness to Action, Springer, Milano, 2010, pp. 145–155, http://dx.doi.org/10.1007/978-88-470-1587-6_8, URL https://link.springer.com/chapter/10.1007/978-88-470-1587-6_8.

[13] J. Moore, P. Fletcher, Sense of agency in health and disease: a review of cue integration approaches, Conscious. Cogn. 21 (1) (2012) 59–68, http://dx.doi.org/10.1016/j.concog.2011.08.010, URL https://www.sciencedirect.com/science/article/pii/S1053810011002005.

[14] T. Zalla, M. Sperduti, The sense of agency in autism spectrum disorders: a dissociation between prospective and retrospective mechanisms? Front. Psych. 6 (2015) http://dx.doi.org/10.3389/fpsyg.2015.01278, URL https://www.frontiersin.org/article/10.3389/fpsyg.2015.01278.

[15] F. Garbarini, A. Mastropasqua, M. Sigaudo, M. Rabuffetti, A. Piedimonte, L. Pia, P. Rocca, Abnormal sense of agency in patients with schizophrenia: evidence from bimanual coupling paradigm, Front. Behav. Neurosci. 10 (2016) http://dx.doi.org/10.3389/fnbeh.2016.00043, URL https://www.frontiersin.org/articles/10.3389/fnbeh.2016.00043/full.

[16] E. Kozáková, E. Bakštein, O. Havlıček, O. Bečev, P. Knytl, Y. Zaytseva, F. Španiel, Disrupted sense of agency as a state marker of first-episode schizophrenia: a large-scale follow-up study, Front. Psychiatry 11 (2020) 395–401, http://dx.doi.org/10.3389/fpsyt.2020.570570, URL https://www.frontiersin.org/articles/10.3389/fpsyt.2020.570570/full.

[17] P. Salgado-Pineda, P. Fuentes-Claramonte, B. Spanlang, A. Pomes, R. Landin-Romero, F. Portillo, C. Bosque, J.C. Franquelo, C. Teixido, S. Sarró, R. Salvador, E.P.-C. Sahaï, N. Hamidi, E. Pacherie, B. Berberian, L. Roche, L. Saint-Bauzel, Neural correlates of disturbance in the sense of agency in schizophrenia: an fMRI study using the 'enfacement' paradigm, Schizophr. Res. 243 (2022) 395–401, http://dx.doi.org/10.1016/j.schres.2021.06.031, URL https://www.sciencedirect.com/science/article/pii/S0920996421002437.

[18] A.R. Krugwasser, Y. Stern, N. Faivre, E.V. Harel, R. Salomon, Impaired sense of agency and associated confidence in psychosis, Schizophr 8 (2022) http://dx.doi.org/10.1038/s41537-022-00212-4, URL https://www.nature.com/articles/s41537-022-00212-4.

[19] M. Synofzik, P. Thier, D.T. Leube, P. Schlotterbeck, A. Lindner, Misattributions of agency in schizophrenia are based on imprecise predictions about the sensory consequences of one's actions, Brain 133 (2010) 262–271, http://dx.doi.org/10.1093/brain/awp291, URL https://academic.oup.com/brain/article/133/1/262/312440.

[20] T.D. Cannon, How schizophrenia develops: cognitive and brain mechanisms underlying onset of psychosis, Trends Cogn. Sci. 19 (2015) 744–756, http://dx.doi.org/10.1016/j.tics.2015.09.009, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4673025/.

[21] S.S. Obhi, P. Hall, Sense of agency in joint action: influence of human and computer co-actors, Exp. Brain Res. 211 (2011) 663–670, http://dx.doi.org/10.1007/s00221-011-2662-7, URL https://link.springer.com/article/10.1007/s00221-011-2662-7.

[22] H. Limerick, D. Coyle, J.W. Moore, The experience of agency in human-computer interactions: a review, Front. Hum. Neurosci. 8 (2014) 643, http://dx.doi.org/10.3389/fnhum.2014.00643.

[23] B. Berberian, J.-C. Sarrazin, P.L. Blaye, P. Haggard, Automation technology and sense of control: a window on human agency, PLoS One 7 (3) (2012) e34075, http://dx.doi.org/10.1371/journal.pone.0034075.

[24] J.E. McEneaney, Agency effects in human–computer interaction, Int. J. Hum.-Comput. Int. 29 (2013) 798–813, http://dx.doi.org/10.1080/10447318.2013.777826, URL https://www.tandfonline.com/doi/abs/10.1080/10447318.2013.777826.

[25] B. Berberian, Man-machine teaming: a problem of agency, IFAC-PapersOnLine 51 (2019) 118–123, http://dx.doi.org/10.1016/j.ifacol.2019.01.049, URL https://www.sciencedirect.com/science/article/pii/S2405896319300515.

[26] B. Shneiderman, C. Plaisant, Designing the User Interface: Strategies for Effective Human-Computer Interaction, fourth ed., Addison Wesley, Reading, MA, 2004, URL https://www.amazon.co.jp/Designing-User-Interface-Human-Computer-Interaction/dp/0321197860.

[27] D. Coyle, J. Moore, P.O. Kristensson, P. Fletcher, A. Blackwell, I did that! measuring users' experience of agency in their own actions, in: Proc. SIGCHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, 2012, pp. 2025–2034, URL https://doi.org/10.1145/2207676.2208350.

[28] J. Bergstrom-Lehtovirta, D. Coyle, J. Knibbe, K. Hornbæk, I really did that: sense of agency with touchpad, keyboard, and on-skin interaction, in: Proc. 2018 CHI Conf. Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, 2018, pp. 1–8, URL https://doi.org/10.1145/3173574.3173952.

[29] K. Le Goff, A. Rey, P. Haggard, O. Oullier, B. Berberian, Agency modulates interactions with automation technologies, Ergonomics 61 (2018) 1282–1297, http://dx.doi.org/10.1080/00140139.2018.1468493.

[30] W. Wen, Y. Kuroki, H. Asama, The sense of agency in driving automation, Front. Psychol. 10 (2019) 2691, http://dx.doi.org/10.3389/fpsyg.2019.02691, URL https://www.frontiersin.org/article/10.3389/fpsyg.2019.02691.

[31] D. Zanatto, M. Chattington, J. Noyes, Sense of agency in human-machine interaction, in: H. Ayaz, U. Asgher, L. Paletta (Eds.), Advances in Neuroergonomics and Cognitive Engineering (AHFE 2021), in: Lecture Notes in Networks and Systems, vol. 259, Springer, Cham, 2021, pp. 353–360, http://dx.doi.org/10.1007/978-3-030-80285-1_41, URL https://link.springer.com/chapter/10.1007/978-3-030-80285-1_41.

[32] W. Wen, H. Imamizu, The sense of agency in perception, behaviour and human–machine interactions, Nat. Rev. Psychol. 1 (2022) 211–222, http://dx.doi.org/10.1038/s44159-022-00030-6, URL https://www.nature.com/articles/s44159-022-00030-6.

[33] Q. Vantrepotte, B. Berberian, M. Pagliari, V. Chambon, Leveraging human agency to improve confidence and acceptability in human-machine interactions, Cogn. 222 (2022) 105020, http://dx.doi.org/10.1016/j.cognition.2022.105020, URL https://www.sciencedirect.com/science/article/pii/S0010027722000087.

[34] M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino, C. Yoshida, Cognitive developmental robotics: A survey, IEEE Trans. Auton. Mental Dev. 1 (1) (2009) 12–34, http://dx.doi.org/10.1109/TAMD.2009.2021702.

[35] J. Tani, An interpretation of the "self" from the dynamical systems perspective: a constructivist approach, J. Conscious. Stud. 5 (1998) 516–542.

[36] J. Tani, Autonomy of self at criticality: the perspective from synthetic neuro-robotics, Adapt. Behav. 17 (5) (2009) 421–443, arXiv:10.1177/1059712309344421. URL https://doi.org/10.1177/1059712309344421.

[37] W. Ohata, J. Tani, Investigation of the sense of agency in social cognition, based on frameworks of predictive coding and active inference: a simulation study on multimodal imitative interaction, Front. Neurorobot. 14 (2020) 61, http://dx.doi.org/10.3389/fnbot.2020.00061, URL https://www.frontiersin.org/article/10.3389/fnbot.2020.00061.

[38] G. Schillaci, C.-N. Ritter, V.V. Hafner, B. Lara, Body representations for robot ego-noise modelling and prediction. towards the development of a sense of agency in artificial agents, Artif. (2016) 390–397, http://dx.doi.org/10.1162/978-0-262-33936-0-ch065, URL https://direct.mit.edu/isal/proceedings/alif2016/28/390/99437.

[39] G. Schillaci, V.V. Hafner, B. Lara, Exploration behaviors, body representations, and simulation processes for the development of cognition in artificial agents, Front. Robot. AI 3 (2016) 39, http://dx.doi.org/10.3389/frobt.2016.00039, URL https://www.frontiersin.org/article/10.3389/frobt.2016.00039.

[40] S. Bechtle, G. Schillaci, V.V. Hafner, On the sense of agency and of object permanence in robots, in: Proc. 2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob), 2016, pp. 166–171, http://dx.doi.org/10.1109/DEVLRN.2016.7846812.

[41] C. Lang, G. Schillaci, V.V. Hafner, A deep convolutional neural network model for sense of agency and object permanence in robots, in: Proc. Joint IEEE 8th Int Conf Development and Learning and Epigenetic Robotics (ICDL-EpiRob), IEEE, 2018, pp. 257–262, http://dx.doi.org/10.1109/DEVLRN.2018.8761015, URL https://ieeexplore.ieee.org/abstract/document/8761015.

[42] C. Lang, G. Schillaci, V.V. Hafner, A deep convolutional neural network model for sense of agency and object permanence in robots, in: Proic. 2018 Joint IEEE 8th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob), 2018, pp. 257–262, http://dx.doi.org/10.1109/DEVLRN.2018.8761015.

[43] R. Legaspi, Z. He, T. Toyoizumi, Synthetic agency: sense of agency in artificial intelligence, Curr. Opin. Behav. Sci. 29 (2019) 84–90, http://dx.doi.org/10.1016/j.cobeha.2019.04.004, Artificial Intelligence. URL https://www.sciencedirect.com/science/article/pii/S2352154618301700.

[44] R.P. van der Wel, Me and we: Metacognition and performance evaluation of joint actions, Cogn. 140 (2015) 49–59, http://dx.doi.org/10.1016/j.cognition.2015.03.011, URL https://www.sciencedirect.com/science/article/pii/S001002771500061Xb.

[45] N.K. Bolt, E.M. Poncelet, B.G. Schultz, J.D. Loehr, Mutual coordination strengthens the sense of joint agency in cooperative joint action, Conscious. Cogn. 46 (2016) 173–187, http://dx.doi.org/10.1016/j.concog.2016.10.001, URL https://www.sciencedirect.com/science/article/abs/pii/S1053810016301398?via%3Dihub.

[46] E.A. Caspar, A. Cleeremans, P. Haggard, Only giving orders? An experimental study of the sense of agency when giving or receiving command, PLoS One 13 (9) (2018) e0204027, http://dx.doi.org/10.1371/journal.pone.0204027, URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0204027.

[47] J.D. Loehr, The sense of agency in joint action: an integrative review, Psychol. Bull. Rev. 29 (2022) 1089–1117, http://dx.doi.org/10.3758/s13423-021-02051-3, URL https://link.springer.com/article/10.3758/s13423-021-02051-32.

[48] T.W. Victor, E. Tivesten, P. Gustavsson, J. Johansson, F. Sangberg, M.L. Aust, Automation expectation mismatch: incorrect prediction despite eyes on threat and hands on wheel, Hum. Factors 60 (2018) 1095–1116, http://dx.doi.org/10.1177/0018720818788164, URL https://journals.sagepub.com/doi/full/10.1177/0018720818788164.

[49] M. Pagliari, V. Chambon, B. Berberian, What is new with Artificial Intelligence? Human-agent interactions through the lens of social agency, Front. Psychol. 13 (2022) 954444, http://dx.doi.org/10.3389/fpsyg.2022.954444, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9559368/.

[50] W. Wen, A. Yamashita, H. Asama, Divided attention and processes underlying sense of agency, Front. Psychol. 7 (2016) http://dx.doi.org/10.3389/fpsyg.2016.00035, URL https://www.frontiersin.org/articles/10.3389/fpsyg.2016.00035.

[51] N. Hon, Attention and the sense of agency: a review and some thoughts on the matter, Conscious. Cogn. 56 (2017) 30–36, http://dx.doi.org/10.1016/j.concog.2017.10.004, URL https://www.sciencedirect.com/science/article/pii/S1053810017303227.

[52] N. Sidarus, P. Haggard, Difficult action decisions reduce the sense of agency: a study using the Eriksen flanker task, Acta Psychol. 166 (2016) 1–11, http://dx.doi.org/10.1016/j.actpsy.2016.03.003, URL https://www.sciencedirect.com/science/article/pii/S0001691816300476.

[53] N. Sidarus, M. Vuorre, J. Metcalfe, P. Haggard, Investigating the prospective sense of agency: effects of processing fluency, stimulus ambiguity, and response conflict, Front. Psychol. 8 (2017) 545, http://dx.doi.org/10.3389/fpsyg.2017.00545, URL https://www.frontiersin.org/article/10.3389/fpsyg.2017.00545.

[54] M. Synofzik, G. Vosgerau, M. Voss, The experience of agency: an interplay between prediction and postdiction, Front. Psychol. 4 (2013) 127, http://dx.doi.org/10.3389/fpsyg.2013.00127, URL https://www.frontiersin.org/article/10.3389/fpsyg.2013.00127.

[55] M. Liesner, W. Kirsch, W. Kunde, The interplay of predictive and postdictive components of experienced selfhood, Conscious. Cogn. 77 (2020) 102850, http://dx.doi.org/10.1016/j.concog.2019.102850, URL https://www.sciencedirect.com/science/article/pii/S1053810019301588.

[56] N. Sidarus, M. Vuorre, P. Haggard, Integrating prospective and retrospective cues to the sense of agency: a multi-study investigation, Neurosci. Conscious. 2017 (2017) 545, http://dx.doi.org/10.1093/nc/nix012, URL https://academic.oup.com/nc/article/2017/1/nix012/3858560.

[57] C.D. Frith, S.-J. Blakemore, D.M. Wolpert, Abnormalities in the awareness and control of action, Philos. Trans. R. Soc. Lond. B. Biol. Sci. 355 (1404) (2000) 1771–1788, http://dx.doi.org/10.1098/rstb.2000.0734, arXiv:https://royalsocietypublishing.org/doi/pdf/10.1098/rstb.2000.0734. URL https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2000.0734.

[58] D.M. Wegner, T. Wheatley, Apparent mental causation: sources of the experience of will, Am. Psychol. 54 (1999) 480–492, http://dx.doi.org/10.1037//0003-066x.54.7.480, URL https://doi.apa.org/doiLanding?doi=10.1037%2F0003-066X.54.7.480.

[59] M. Synofzik, G. Vosgerau, A. Newen, Beyond the comparator model: a multifactorial two-step account of agency, Conscious. Cog. 17 (1) (2008) 219–239, http://dx.doi.org/10.1016/j.concog.2007.03.010, URL https://www.sciencedirect.com/science/article/pii/S1053810007000268.

[60] S.J. Blakemore, D.M. Wolpert, C.D. Frith, Affectivity and the distinction between minimal and narrative self, Trends Cogn. Sci. 6 (2002) 237–242, http://dx.doi.org/10.1016/s1364-6613(02)01907-1, URL https://pubmed.ncbi.nlm.nih.gov/12039604/.

[61] D.M. Wegner, The Illusion of Conscious Will, MIT Press, Cambridge, 2002, p. 440, URL https://mitpress.mit.edu/books/illusion-conscious-will.

[62] D.M. Wegner, The mind's best trick: How we experience conscious will, Trends Cogn. Sci. 7 (2003) 65–69, http://dx.doi.org/10.1016/S1364-6613(03)00002-0, URL https://www.sciencedirect.com/science/article/pii/S1364661303000020.

[63] N. Braun, S. Debener, N. Spychala, E. Bongartz, P. Sörös, H.H.O. Müller, A. Philipsen, The senses of agency and ownership: a review, Front. Psychol. 9 (2018) 535, http://dx.doi.org/10.3389/fpsyg.2018.00535, URL https://www.frontiersin.org/article/10.3389/fpsyg.2018.00535.

[64] R. Legaspi, T. Toyoizumi, A Bayesian psychophysics model of sense of agency, Nature Commun. 10 (2019) 4250, http://dx.doi.org/10.1038/s41467-019-12170-0, URL https://www.nature.com/articles/s41467-019-12170-0.

[65] P. Haggard, S. Clark, J. Kalogeras, Voluntary action and conscious awareness, Nature Neurosci. 5 (2002) 382–385, http://dx.doi.org/10.1038/nn827, URL https://doi.org/10.1038/nn827.

[66] N. Wolpe, P. Haggard, H.R. Siebner, J.B. Rowe, Cue integration and the perception of action in intentional binding, Exp. Brain Res. 229 (2013) 467–474, http://dx.doi.org/10.1007/s00221-013-3419-2.x, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3745826/.

[67] G. Carruthers, What makes us conscious of our own agency? And why the conscious versus unconscious representation distinction matter, Front. Hum. Neurosci. 8 (2014) 434, http://dx.doi.org/10.3389/fnhum.2014.00434, URL https://www.frontiersin.org/article/10.3389/fnhum.2014.00434.

[68] P. Boyer, P. Robbins, A.I. Jack, Varieties of self-systems worth having, Conscious. Cogn. 14 (4) (2005) 647–660, http://dx.doi.org/10.1016/j.concog.2005.08.002, URL https://www.sciencedirect.com/science/article/pii/S1053810005001169.

[69] S.-J. Blakemore, D.M. Wolpert, C.D. Frith, Central cancellation of self-produced tickle sensation, Nature Neurosci. 1 (1998) 635–640, http://dx.doi.org/10.1038/2870, URL https://www.nature.com/articles/nn1198_635.

[70] S.-J. Blakemore, D. Wolpert, C. Frith, Why can't you tickle yourself? NeuroReport 11 (2000) R11–R16, http://dx.doi.org/10.1097/00001756-200008030-00002, URL https://journals.lww.com/neuroreport/Fulltext/2000/08030/Why_can_t_you_tickle_yourself_.2.aspx.

[71] J.W. Moore, S.S. Obhi, Intentional binding and the sense of agency: a review, Conscious. Cogn. 21 (2012) 546–561, http://dx.doi.org/10.1016/j.concog.2011.12.002, URL https://www.sciencedirect.com/science/article/pii/S1053810011002881.

[72] E. Hildt, Artificial intelligence: does consciousness matter? Front. Psychol. 10 (2019) 1535, http://dx.doi.org/10.3389/fpsyg.2019.01535, URL https://www.frontiersin.org/article/10.3389/fpsyg.2019.01535.

[73] R. Chatila, E. Renaudo, M. Andries, R.-O. Chavez-Garcia, P. Luce-Vayrac, R. Gottstein, R. Alami, A. Clodic, S. Devin, B. Girard, M. Khamassi, Toward self-aware robots, Front. Robot. AI 5 (2018) 88, http://dx.doi.org/10.3389/frobt.2018.00088, URL https://www.frontiersin.org/article/10.3389/frobt.2018.00088.

[74] V.V. Hafner, P. Loviken, A. Pico Villalpando, G. Schillaci, Prerequisites for an artificial self, Front. Neurorobot. 14 (2020) 5, http://dx.doi.org/10.3389/fnbot.2020.00005, URL https://www.frontiersin.org/article/10.3389/fnbot.2020.00005.

[75] J.J. Bryson, M.E. Diamantis, T.D. Grant, Of, for, and by the people: the legal lacuna of synthetic persons, Artif. Intell. Law 25 (2017) 273–291, http://dx.doi.org/10.1007/s10506-017-9214-9, URL https://doi.org/10.1007/s10506-017-9214-9.

[76] S.M. Solaiman, Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy, Artif. Intell. Law 25 (2017) 155–179, http://dx.doi.org/10.1007/s10506-016-9192-3, URL https://doi.org/10.1007/s10506-016-9192-3.

[77] P.J. Hinds, User Control and its Many Facets: a Study of Perceived Control in Human-Computer Interaction, Tech. Rep. HPL-98-154, Hewlett Packard, HP Laboratories, California, 1998.

[78] G.A. Zachiotis, G. Andrikopoulos, R. Gornez, K. Nakamura, G. Nikolakopoulos, A survey on the application trends of home service robotics, in: Proc. 2018 IEEE Int. Conf. Robotics and Biomimetics, ROBIO, IEEE, 2018, pp. 1999–2006, http://dx.doi.org/10.1109/ROBIO.2018.8665127, URL https://ieeexplore.ieee.org/document/8665127.

[79] W.D. Stiehl, C. Breazeal, J. Tao, T. Tan, R.W. Picard, Affective touch for robotic companions, in: Affective Computing and Intelligent Interaction. ACII 2005, in: Lecture Notes in Computer Science, vol. 3784, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 747–754, http://dx.doi.org/10.1007/11573548_96, URL https://link.springer.com/chapter/10.1007/11573548_96.

[80] L. Pu, W. Moyle, C. Jones, M. Todorovic, The effectiveness of social robots for older adults: a systematic review and meta-analysis of randomized controlled studies, Gerontology 59 (2019) e37–e51, http://dx.doi.org/10.1093/geront/gny046, URL https://academic.oup.com/gerontologist/article/59/1/e37/5036100.

[81] S. Satake, K. Nakatani, K. Hayashi, T. Kanda, M. Imai, What should we know to develop an information robot? PeerJ Comput. Sci. 1:e8 (2015) http://dx.doi.org/10.7717/peerj-cs.8, URL https://peerj.com/articles/cs-8/.

[82] L. Van Aerschot, J. Parviainen, Robots responding to care needs? A multitasking care robot pursued for 25 years, available products offer simple entertainment and instrumental assistance, Ethics Inf. Technol. 22 (2020) 247–256, http://dx.doi.org/10.1007/s10676-020-09536-0, URL https://link.springer.com/article/10.1007/s10676-020-09536-0.

[83] E. Broadbent, Interactions with robots: the truths we reveal about ourselves, Annu. Rev. Psychol. 68 (2017) 627–652, http://dx.doi.org/10.1146/annurev-psych-010416-043958, URL https://www.annualreviews.org/doi/10.1146/annurev-psych-010416-043958.

[84] A. Fast-Berglund, P. Thorvald, E. Billing, A. Palmquist, D. Romero, G. Weichhart, Conceptualizing embodied automation to increase transfer of tacit knowledge in the learning factory, in: Proc. 2018 Int. Conf. Intelligent Systems, IS, 2018, pp. 358–364, http://dx.doi.org/10.1109/IS.2018.8710482, URL https://ieeexplore.ieee.org/document/8710482.

[85] F. Ciardo, D. De Tommaso, F. Beyer, A. Wykowska, Reduced sense of agency in human-robot interaction, in: S.S. Ge, J.-J. Cabibihan, M.A. Salichs, E. Broadbent, H. He, A.R. Wagner, A. Castro-González (Eds.), Social Robotics. ICSR 2018, in: Lecture Notes in Computer Science, vol. 11357, Springer International Publishing, Cham, 2018, pp. 441–450, http://dx.doi.org/10.1007/978-3-030-05204-1_43, URL https://link.springer.com/chapter/10.1007/978-3-030-05204-1_43.

[86] A. Sahaï, A. Desantis, O. Grynszpan, E. Pacherie, B. Berberian, Action co-representation and the sense of agency during a joint simon task: comparing human and machine co-agents, Conscious. Cogn. 67 (2019) 44–55, http://dx.doi.org/10.1016/j.concog.2018.11.008, URL https://www.sciencedirect.com/science/article/pii/S1053810018301156.

[87] Z. Barlas, When robots tell you what to do: sense of agency in human- and robot-guided actions, Conscious. Cogn. 75 (2019) 102819, http://dx.doi.org/10.1016/j.concog.2019.102819, URL https://www.sciencedirect.com/science/article/pii/S1053810019301850.

[88] C. Nass, J. Steuer, E.R. Tauber, Computers are social actors, in: Proc. SIGCHI Conf. Human Factors in Computing Systems, CHI '94, Association for Computing Machinery, New York, NY, USA, 1994, pp. 72–78, http://dx.doi.org/10.1145/191666.191703, URL https://dl.acm.org/doi/10.1145/191666.191703.

[89] K. Nagao, A. Takeuchi, Social interaction: multimodal conversation with social agents, in: Proc. Twelfth National Conf. Artificial Intelligence (Vol. 1), AAAI '94, American Association for Artificial Intelligence, USA, 1994, pp. 22–28.

[90] D.C. Dennett, The Intentional Stance, MIT Press, Cambridge, MA, 1987.

[91] J. Złotowski, D. Proudfoot, K. Yogeeswaran, C. Bartneck, Anthropomorphism: opportunities and challenges in human–robot interaction, Int. J. Soc. Robot. 7 (2015) 347–360, http://dx.doi.org/10.1007/s12369-014-0267-6, URL https://link.springer.com/article/10.1007%2Fs12369-014-0267-6.

[92] F. Ciardo, F. Beyer, D. De Tommaso, A. Wykowska, Attribution of intentional agency towards robots reduces one's own sense of agency, Cogn. 194 (2020) 104109, http://dx.doi.org/10.1016/j.cognition.2019.104109, URL https://www.sciencedirect.com/science/article/pii/S0010027719302835.

[93] C. Roselli, F. Ciardo, A. Wykowska, Intentions with actions: the role of intentionality attribution on the vicarious sense of agency in Human-Robot interaction, Q. J. Exp. Psychol. (Hove) 75 (2022) 616–632, http://dx.doi.org/10.1177/17470218211042003, URL https://journals.sagepub.com/doi/10.1177/17470218211042003.

[94] C. Roselli, F. Ciardo, D. De Tommaso, A. Wykowska, Human-likeness and attribution of intentionality predict vicarious sense of agency over humanoid robot actions, Sci. Rep. 12 (2022) 13845, http://dx.doi.org/10.1038/s41598-022-18151-6, URL https://www.nature.com/articles/s41598-022-18151-6.

[95] R. Pfeifer, F. Iida, Embodied artificial intelligence: trends and challenges, in: F. Iida, R. Pfeifer, L. Steels, Y. Kuniyoshi (Eds.), Embodied Artificial Intelligence, in: Lecture Notes in Computer Science, vol. 3139, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004, pp. 1–26, http://dx.doi.org/10.1007/978-3-540-27833-7_1, URL https://link.springer.com/chapter/10.1007/978-3-540-27833-7_1.

[96] X.E. Barandiaran, E.D. Paolo, M. Rohde, Defining agency: individuality, normativity, asymmetry, and spatio-temporality in action, Adapt. Behav. 17 (5) (2009) 367–386, arXiv:10.1177/1059712309343819. URL https://doi.org/10.1177/1059712309343819.

[97] K. Dautenhahn, B. Ogden, T. Quick, From embodied to socially embedded agents — implications for interaction-aware robots, Cogn. Syst. Res. 3 (2002) 397–428, http://dx.doi.org/10.1016/S1389-0417(02)00050-5, URL https://www.sciencedirect.com/science/article/pii/S1389041702000505.

[98] P. Milgram, F. Kishino, A taxonomy of mixed reality visual displays, IEICE Trans. Inf. & Syst. E77-D (1994) 1321–1329, URL https://search.ieice.org/bin/summary.php?id=e77-d_12_1321.

[99] R.J. Beun, E.D. Vos, C. Witteman, Embodied conversational agents: effects on memory performance and anthropomorphisation, in: T. Rist, R.S. Aylett, D. Ballin, J. Rickel (Eds.), Proc. 4th Int. Workshop on Intelligent Virtual Agents, in: Lecture Notes in Computer Science, vol. 2792, Springer Berlin Heidelberg, Berlin, Heidelberg, 2003, pp. 315–319, http://dx.doi.org/10.1007/978-3-540-39396-2_52, URL https://link.springer.com/chapter/10.1007%2F978-3-540-39396-2_52.

[100] Y. Jie, J.Y. Pei, L. Jun, G. Yun, X. Wei, Smart home system based on IOT technologies, in: International Conference on Computational and Information Sciences, IEEE, 2013, pp. 1789–1791, http://dx.doi.org/10.1109/ICCIS.2013.468, URL https://ieeexplore.ieee.org/document/6643387.

[101] E. Okai, X. Feng, P. Sant, Smart cities survey, in: 20th International Conference on High Performance Computing and Communications; 16th International Conference on Smart City; 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), IEEE, 2018, pp. 1726–1730, http://dx.doi.org/10.1109/HPCC/SmartCity/DSS.2018.00282, URL https://ieeexplore.ieee.org/document/8623018.

[102] D. Panzoli, C. Peters, I. Dunwell, S. Sanchez, P. Petridis, A. Protopsaltis, V. Scesa, S. de Freitas, A level of interaction framework for exploratory learning with characters in virtual environments, in: D. Plemenos, G. Miaoulis (Eds.), Intelligent Computer Graphics 2010, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 123–143, http://dx.doi.org/10.1007/978-3-642-15690-8_7, URL https://link.springer.com/chapter/10.1007/978-3-642-15690-8_7.

[103] S. Zafari, S.T. Koeszegi, Attitudes toward attributed agency: role of perceived control, Int. J. Soc. Robot. (2020) http://dx.doi.org/10.1007/s12369-020-00672-7, URL https://link.springer.com/article/10.1007%2Fs12369-020-00672-7.

[104] F. Beyer, N. Sidarus, S. Bonicalzi, P. Haggard, Beyond self-serving bias: diffusion of responsibility reduces sense of agency and outcome monitoring, Soc. Cogn. Affect. Neurosci. 12 (2017) 138–145, http://dx.doi.org/10.1093/scan/nsw160, URL https://academic.oup.com/scan/article/12/1/138/2628052.

[105] F. Beyer, N. Sidarus, S. Fleming, P. Haggard, Losing control in social situations: how the presence of others affects neural processes related to sense of agency, eNeuro 5 (2018) http://dx.doi.org/10.1523/ENEURO.0336-17.2018, URL https://www.eneuro.org/content/5/1/ENEURO.0336-17.2018.

[106] P. Haggard, M. Tsakiris, The experience of agency: feelings, judgments, and responsibility, Curr. Dir. Psychol. Sci. 18 (4) (2009) 242–246, http://dx.doi.org/10.1111/j.1467-8721.2009.01644.x, URL https://journals.sagepub.com/doi/10.1111/j.1467-8721.2009.01644.x.

[107] L. Takayama, Perspectives on agency interacting with and through personal robots, in: M. Zacarias, J.V. de Oliveira (Eds.), Human-Computer Interaction: The Agency Perspective, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 195–214, http://dx.doi.org/10.1007/978-3-642-25691-2_8, URL https://link.springer.com/chapter/10.1007/978-3-642-25691-2_8.

[108] M.K. Hui, J.E.G. Bateson, Perceived control and the effects of crowding and consumer choice on the service experience, J. Consum. Res. 18 (1991) 174–184, URL http://www.jstor.org/stable/2489553.

[109] J.C. Ward, J.W. Barnes, Control and affect: the influence of feeling in control of the retail environment on affect, involvement, attitude, and behavior, J. Bus. Res. 54 (2001) 139–144, http://dx.doi.org/10.1016/S0148-2963(99)00083-1, URL https://www.sciencedirect.com/science/article/abs/pii/S0148296399000831.

[110] C.-C. Chang, Choice, perceived control, and customer satisfaction: the psychology of online service recovery, Cyberpsychol. Behav. 11 (2008) 321–328, http://dx.doi.org/10.1089/cpb.2007.0059, URL https://www.liebertpub.com/doi/10.1089/cpb.2007.0059.

[111] B.M. Noone, Customer perceived control and the moderating effect of restaurant type on evaluations of restaurant employee performances, Int. J. Hosp. Manag. 27 (2008) 23–29, http://dx.doi.org/10.1016/j.ijhm.2007.07.002, URL https://www.sciencedirect.com/science/article/pii/S0278431907000382.

[112] B.M. Noone, J. Wirtz, S.E. Kimes, The effect of perceived control on consumer responses to service encounter pace: a revenue management perspective, Cornell Hosp. Q. 53 (2012) 295–307, http://dx.doi.org/10.1177/1938965512460343, URL https://journals.sagepub.com/doi/10.1177/1938965512460343.

[113] N.A. Pacheco, R. Lunardo, C.P.d. Santos, A perceived-control based model to understanding the effects of co-production on satisfaction, BAR - Braz. Adm. Rev. 10 (2013) 219–238, http://dx.doi.org/10.1590/S1807-76922013000200007, URL https://www.scielo.br/j/bar/a/VZhG4dbkRHWKCP4C6QGSKpc/?lang=en#.

[114] A. Faraji-Rad, S. Melumad, G.V. Johar, Consumer desire for control as a barrier to new product adoption, J. Consum. Psychol. 27 (2017) 347–354, http://dx.doi.org/10.1016/j.jcps.2016.08.002, URL https://myscp.onlinelibrary.wiley.com/doi/abs/10.1016/j.jcps.2016.08.002.

[115] D.A. Norman, The 'problem' with automation: inappropriate feedback and interaction, not 'over-automation', Philos. Trans. R. Soc. Lond. B Biol. Sci. 327 (1990) 585–593, http://dx.doi.org/10.1098/rstb.1990.0101, URL https://royalsocietypublishing.org/doi/epdf/10.1098/rstb.1990.0101.

[116] C. Rudin, J. Radin, Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition, Harvard Data Sci. Rev. 1 (2019) http://dx.doi.org/10.1162/99608f92.5a8a3a3d, URL https://hdsr.mitpress.mit.edu/pub/f9kuryi8/release/8.

[117] V. Buhrmester, D. Münch, M. Arens, Analysis of explainers of black box deep neural networks for computer vision: a survey, 2019, arXiv:1911.12116.

[118] A.F. Markus, J.A. Kors, P.R. Rijnbeek, The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies, J. Biomed. Inform. 113 (2021) 103655, http://dx.doi.org/10.1016/j.jbi.2020.103655, URL https://www.sciencedirect.com/science/article/pii/S1532046420302835.

[119] A. Rawal, J. McCoy, D. Rawat, B. Sadler, R. Amant, Recent advances in trustworthy explainable artificial intelligence: status, challenges and perspectives, 2021, http://dx.doi.org/10.36227/techrxiv.17054396.v1.

[120] A. Tapal, E. Oren, R. Dar, B. Eitam, The sense of agency scale: a measure of consciously perceived control over one's mnd, body, and the immediate environment, Front. Psychol. 8 (2017) 1552, http://dx.doi.org/10.3389/fpsyg.2017.01552, URL https://www.frontiersin.org/article/10.3389/fpsyg.2017.01552.

[121] K. Borhani, B. Beck, P. Haggard, Choosing, doing, and controlling: implicit sense of agency over somatosensory events, Psychol. Sci. 28 (2017) 882–893, http://dx.doi.org/10.1177/0956797617697693, URL https://journals.sagepub.com/doi/abs/10.1177/0956797617697693.

[122] R. Legaspi, W. Xu, T. Konishi, S. Wada, Positing a sense of agency-aware persuasive AI: its theoretical and computational frameworks, in: R. Ali, B. Lugrin, F. Charles (Eds.), Persuasive Technology: 16th Int. Conf. PERSUASIVE 2021, in: Information Systems and Applications, incl. Internet/Web, and HCI, vol. 12684, Springer International Publishing, 2021, p. XII, 330, http://dx.doi.org/10.1007/978-3-030-79460-6.

[123] R. Legaspi, W. Xu, T. Konishi, S. Wada, Y. Ishikawa, Multidimensional analysis of sense of agency during goal pursuit, in: Proc. 30th ACM Conf. User Modeling, Adaptation and Personalization, UMAP'22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 34–47, http://dx.doi.org/10.1145/3503252.3531303, URL https://dl.acm.org/doi/10.1145/3503252.3531303.

[124] A. Gentsch, M. Synofzik, Affective coding: the emotional dimension of agency, Fronti. Hum. Neurosci. 8 (2014) 608, http://dx.doi.org/10.3389/fnhum.2014.00608, URL https://www.frontiersin.org/article/10.3389/fnhum.2014.00608.

[125] J. Tao, T. Tan, Affective computing: a review, in: J. Tao, T. Tan, R.W. Picard (Eds.), Affective Computing and Intelligent Interaction, in: Lecture Notes in Computer Science, vol. 3784, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 981–995, http://dx.doi.org/10.1007/11573548_125, URL https://link.springer.com/chapter/10.1007/11573548_125.

[126] J. Pearl, D. Mackenzie, The Book of Why: The New Science of Cause and Effect, Basic Books, 2018.

[127] E. Bareinboim, J. Pearl, Causal inference and the data-fusion problem, Proc. Natl. Acad. Sci. (PNAS) 113 (2016) 7345–7352, http://dx.doi.org/10.1073/pnas.1510507113, URL https://www.pnas.org/content/113/27/7345.

[128] A.B. Arrieta, N. Díaz-Rodríguez, J.D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115, http://dx.doi.org/10.1016/j.inffus.2019.12.012, URL https://www.sciencedirect.com/science/article/pii/S1566253519308103.

[129] T. Chakraborti, S. Sreedharan, S. Kambhampati, The emerging landscape of explainable automated planning & decision making, in: C. Bessiere (Ed.), Proc. Twenty-Ninth Int. Joint Conf. Artificial Intelligence (IJCAI-20), 2020, pp. 4803–4811, Survey track. URL https://www.ijcai.org/proceedings/2020/669.

[130] W. Xu, Y. Kuriki, T. Sato, M. Taya, C. Ono, Does traffic information provided by smartphones increase detour behavior? in: S.B. Gram-Hansen, T.S. Jonasen, C. Midden (Eds.), Persuasive Technology. Designing for Future Change, Springer International Publishing, Cham, 2020, pp. 45–57, http://dx.doi.org/10.1007/978-3-030-45712-9_4, URL https://link.springer.com/chapter/10.1007/978-3-030-45712-9_4.

[131] R.A. Renes, H. Aarts, The sense of agency in health and well-being: understanding the role of the minimal self in action-control, in: D. de Ridder, M. Adriaanse, K. Fujita (Eds.), The Routledge International Handbook of Self-Control in Health and Well-Being: Concepts, Theories, and Central Issues, Routledge / Taylor and Francis Group, 2017, pp. 193–205, http://dx.doi.org/10.4324/9781315648576, URL https://www.taylorfrancis.com/chapters/edit/10.4324/9781315648576-16/sense-agency-health-well-being-robert-renes-henk-aarts.

[132] H. Aarts, E. Bijleveld, R. Custers, M. Dogge, M. Deelder, D. Schutter, N.E.M. van Haren, Positive priming and intentional binding: eye-blink rate predicts reward information effects on the sense of agency, Soc. Neurosci. 7 (1) (2012) 105–112, http://dx.doi.org/10.1080/17470919.2011.590602, URL https://www.tandfonline.com/doi/abs/10.1080/17470919.2011.590602.

[133] W.C. Willett, M.J. Stampfer, Current evidence on healthy eating, Annu. Rev. Public Health 34 (2013) 77–95, http://dx.doi.org/10.1146/annurev-publhealth-031811-124646, URL https://www.annualreviews.org/doi/10.1146/annurev-publhealth-031811-124646.

[134] C.S. Levine, Y. Miyamoto, H.R. Markus, A. Rigotti, J.M. Boylan, J. Park, S. Kitayama, M. Karasawa, N. Kawakami, C.L. Coe, G.D. Love, C.D. Ryff, Culture and healthy eating: the role of independence and interdependence in the United States and Japan, Pers. Soc. Psychol. Bull. 42 (10) (2016) 1335–1348, http://dx.doi.org/10.1177/0146167216658645, URL https://doi.org/10.1177/0146167216658645.

[135] R. Legaspi, Y. Hashimoto, K. Moriyama, S. Kurihara, M. Numao, Music compositional intelligence with an affective flavor, in: Proc 12th Int Conf Intelligent User Interfaces, IUI '07, Association for Computing Machinery, New York, NY, USA, 2007, pp. 216–224, http://dx.doi.org/10.1145/1216295.1216335, URL https://dl.acm.org/doi/abs/10.1145/1216295.1216335.

[136] R.A. Cabredo, R. Legaspi, P.S. Inventado, M. Numao, An emotion model for music using brain waves, in: Proc 13th Int Soc Music Information Retrieval Conf, in: ISMR 2012, 2012, pp. 265–270, URL https://animorepository.dlsu.edu.ph/faculty_research/4438.

[137] L.C. nero Gomez, R. Hervas, I. Gonzalez, L. Rodriguez-Benitez, Eeglib: A python module for EEG feature extraction, SoftwareX 15 (2021) 100745, http://dx.doi.org/10.1016/j.softx.2021.100745, URL https://www.sciencedirect.com/science/article/pii/S2352711021000753.

[138] L. Breiman, Statistical modeling: the two cultures, Stat. Sci. 16 (2001) 199–215, http://dx.doi.org/10.1214/ss/1009213726, URL https://www.jstor.org/stable/2676681.

[139] T.J. Sejnowski, The unreasonable effectiveness of deep learning in artificial intelligence, Proc. Natl. Acad. Sci. 117 (2020) 30033–30038, http://dx.doi.org/10.1073/pnas.1907373117, URL https://www.pnas.org/doi/abs/10.1073/pnas.1907373117.

[140] P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, B.B. Gupta, X. Chen, X. Wang, A survey of deep active learning, ACM Comput. Surv. 54 (2021) 1–40, http://dx.doi.org/10.1145/3472291, URL https://dl.acm.org/doi/abs/10.1145/3472291.

[141] A. Slivkins, Introduction to multi-armed bandits, 2022, arXiv:1904.07272.

Further reading

[1] B. Inden, Z. Malisz, P. Wagner, I. Wachsmuth, Timing and entrainment of multimodal backchanneling behavior for an embodied conversational agent, in: Proc 15th ACM Int Conf Multimodal Interaction, ICMI '13, 2013, pp. 181–188, http://dx.doi.org/10.1145/2522848.2522890.