frontiers in Robotics and AI

HYPOTHESIS AND THEORY
published: 23 April 2021
doi: 10.3389/frobt.2021.620026

Automation Inner Speech as an Anthropomorphic Feature Affecting Human Trust: Current Issues and Future Directions

Alessandro Geraci[1,2], *Antonella D'Amico*[2*], *Arianna Pipitone*[1], *Valeria Seidita*[1] *and Antonio Chella*[1]

[1] *Robotics Lab, Department of Engineering, University of Palermo, Palermo, Italy,* [2] *Department of Psychology, Educational Science and Human Movement, University of Palermo, Palermo, Italy*

This paper aims to discuss the possible role of inner speech in influencing trust in human–automation interaction. Inner speech is an everyday covert inner monolog or dialog with oneself, which is essential for human psychological life and functioning as it is linked to self-regulation and self-awareness. Recently, in the field of machine consciousness, computational models using different forms of robot speech have been developed that make it possible to implement inner speech in robots. As is discussed, robot inner speech could be a new feature affecting human trust by increasing robot transparency and anthropomorphism.

Keywords: inner speech, trust, anthropomorphism, human-automation interaction, human-robot interaction, robot, automation

OPEN ACCESS

Edited by:
Marco Nørskov,
Aarhus University, Denmark

Reviewed by:
James A. Reggia,
University of Maryland, United States
Felix Lindner,
University of Ulm, Germany

***Correspondence:**
Antonella D'Amico
antonella.damico@unipa.it

Specialty section:
This article was submitted to Ethics in Robotics and Artificial Intelligence, a section of the journal Frontiers in Robotics and AI

Received: *21 October 2020*
Accepted: *26 February 2021*
Published: *23 April 2021*

Citation:
Geraci A, D'Amico A, Pipitone A, Seidita V and Chella A (2021) Automation Inner Speech as an Anthropomorphic Feature Affecting Human Trust: Current Issues and Future Directions. Front. Robot. AI 8:620026. doi: 10.3389/frobt.2021.620026

INTRODUCTION

In the past years, robots and automation development and implementation have increased exponentially in every context, leading to growing interactions with humans (Merritt and Ilgen, 2008). Robots are now used in different contexts, such as military, security, medical, domestic, and entertainment (Li et al., 2010). Robots, compared with other types of automation (e.g., machines, computers), are designed to be self-governed to some extent to respond to situations that are not prearranged (Lewis et al., 2018). Therefore, the greater the complexity of robots, the higher the importance to focus on factors that influence human–automation interaction (HAI) as their collaboration increases over time (Lee and See, 2004; Schaefer et al., 2016). In this paper, we aim to start the exploration of the role of inner speech in HAI and, in particular, on its role in improving human trust toward automation. For this purpose, we first focus on the concept of inner speech in psychological literature, also examining the first results of its implementation in automation. Then, we discuss the possible role of inner speech as one of the anthropomorphic automation features that may affect human trust in HAI.

INNER SPEECH

Inner speech is an everyday covert inner monolog or dialog with oneself, which is essential for human psychological life and functioning because it is linked to reasoning, self-regulation, and self-awareness (Morin, 2012).

For its nature, inner speech is intrinsically dialogic because it involves the verbal perspective of thoughts. It is not an image or a sensation or pure emotion. Non-verbal thinking is not inner speech.

Inner speech takes the form of a monolog if the communication is one-sided, and it takes the form of dialog when it includes more than one perspective. More specifically, the monolog form involves a conversation with oneself, in which only one point of view is expressed, and an answer is not required (Oleś et al., 2020). On the contrary, the dialog form refers to a simulated exchange between two or more "selves" or between oneself and other imaginary interlocutors, in which two or more points of view or perspectives are taken into account (Fernyhough, 2016).

Nowadays, there are many alternative terms used to refer to inner speech, such as inner voice, private speech, inner language, internal dialog, self-talk, covert speech (Loevenbruck et al., 2018). However, the most accepted definition describes inner speech as "the subjective experience of language in the absence of overt and audible articulation" (Alderson-Day and Fernyhough, 2015, p. 931). Interest in inner speech originates from the psychological literature, particularly from the theoretical debate on the relationship between language and thinking and on the role of inner speech for cognitive development.

Watson (1913), the father of behaviorism, equated inner speech with thinking, affirming that external and inner speech share the same structures except for the articulatory components: child's overt speech transitions to covert speech, passing by whispering, simply through a process of reduction of audible volume. Piaget (1959) named inner speech as egocentric speech, emerging during children's playtime, which he believed to be intimately related to action. He considered egocentric speech to have no specific functions and, thus, to be an egocentric thinking expression. In this early stage, the child cannot discern his perspective from others, destined to disappear, giving way to social language gradually.

Vygotsky (1962), on the contrary, attributed great importance to inner speech as one of the most crucial processes for cognitive and social development. According to Vygotsky (1962), inner speech serves multiple cognitive functions, such as problem solving and self-regulation, because it allows using and controlling thought to plan and monitor behaviors and actions.

Vygotsky (1962) also argued that external and inner speech are almost opposites because "external speech is a process of transforming thought into word; it is the materialization and objectification of thought. Inner speech moves in the reverse direction [...] it is a process that involves the evaporation of speech in thought" (p. 258). He reasoned that early linguistic, social interaction between the child and the caregivers are gradually internalized and transformed into covert self-directed speech. As internalization progresses, the child becomes more psychologically autonomous and self-regulated because "a function becomes internalized when it can be fulfilled without the immediate collaboration of others" (Larrain and Haye, 2012, p. 6).

Despite Vygotsky's fundamental theoretical contribution to inner speech and its central role in human psychological development, over the years, scientific research has shown little attention to this field (Scott et al., 2013). One reason may be due to a general assumption that inner speech follows overt speech form and structure (McCarthy-Jones and Fernyhough, 2011). Another reason is linked to methodological issues in the assessment methods (Alderson-Day and Fernyhough, 2015) because inner speech can be neither observed directly nor behaviorally (Martin et al., 2018), and it can vary in terms of frequency among people (Ren et al., 2016).

More recently, there has been a renewed interest in inner speech: McCarthy-Jones and Fernyhough (2011), following the Vygotsky perspective, argue that inner speech qualitatively differs from overt speech because it has a dialogic and condensed nature; it engages the presence of other people's voices, and it is involved in self and other evaluations. That is because "talking to oneself can instigate a fictional dialog in which [...] people sometimes [...] express to a real or imaginary person their reasons for behaving in a given way or for possessing some personal attributes" (Morin, 2004, pp. 212–213). Inner speech can vary in syntax, semantics, and phonology, spanning from a fully expanded speech to a highly condensed form (Fernyhough, 2004).

The interest in inner speech also depends on recognizing its contribution to other cognitive processes, such as working memory (Baddeley and Hitch, 1974). Working memory is one of the most studied crucial cognitive processes because it does not attain the maintenance of information exclusively but also allows the elaboration of incoming information during a complicated task. For instance, reading, reasoning, or taking a conversation are all processes managed by working memory because, at the same time, we have to hold recent information in short-term memory, to recover old information from long-term memory, and to orient our attention toward the incoming information. Inner speech is considered to support the phonological loop, the working memory slave system responsible for the representation of acoustic, verbal, or phonological information. By inner speech (also referred to as rehearsal process), people maintain recent verbal information in memory while new information is being processed or old information is recovered from long-term memory.

Recently, different studies have demonstrated that inner speech is related to various psychological processes: Tullett and Inzlicht (2010) find that suppressing inner speech, using articulatory suppression tasks, increased impulsive responding during go/no-go tasks, concluding that inner speech is involved in self-control abilities. Gade and Paelecke (2019) show inner speech is linked to conflict resolution abilities and cognitive flexibility as they demonstrate that it improved participants' performances in the Simon task. In addition, studies found that expressing overtly one's mental verbalization during a task (i.e., thinking aloud protocol) facilitates problem-solving and reading comprehension because it helps to be more focused and more capable of following a sequence of self-instruction (Short et al., 1991; Kucan and Beck, 1997).

Inner speech is also associated with self-awareness, a multisource psychological ability to orient attention to oneself (Morin, 2011) because "one becomes self-aware when one

engages in self-talk (higher-order thought) about one's current mental state and personal characteristics" (p. 212). According to Morin (2012), inner speech allows recognizing different self-facets, representing internal states, and consequently thinking about them.

Psychological studies show that inner speech also has a "dark side" because it is involved in different psychopathological disorders (for reviews, see Alderson-Day and Fernyhough, 2015). It is related to auditory verbal hallucinations, a prominent symptom in psychotic disorders, especially schizophrenia (Waters et al., 2012), which refers to the subjective experience of hearing voices in the absence of a speaking source. The most supported theory states that these symptoms derive from a deficit in self-recognition processes. People fail to recognize their thoughts and behaviors as self-generated, consequently believing them to originate from an external source (Frith and Done, 1988; Bentall, 1990; Waters et al., 2012). Other studies show that inner speech is also involved in anxiety and mood disorders: the process of rumination (i.e., repetitive presence of negative thoughts), which is predominantly verbal, generates a negative emotional and cognitive loop maintaining or intensifying the levels of anxiety and depression (McCarthy-Jones and Fernyhough, 2011; Whitmer and Gotlib, 2012; Alderson-Day and Fernyhough, 2015).

Inner speech has been also studied from a neuroscientific perspective. Some studies show that different brain regions activate when inner speech takes the form of either a monolog or a dialog. In the monolog scenario, the inner speech involves the activation of left frontotemporal regions associated with overt speech production and understanding (e.g., left inferior frontal gyrus, middle temporal gyrus: McGuire et al., 1996; Shergill et al., 2002). On the contrary, the dialogic inner speech involves an extensive neural network between two hemispheres (e.g., left and right superior temporal gyri, posterior cingulate: Alderson-Day et al., 2016).

INNER SPEECH AS AN EMERGING AREA OF RESEARCH IN AI

Inner speech is also an emerging area of interest in the field of artificial intelligence. Over the last two decades, various computational models have included simulations of different inner speech forms (Reggia, 2013). For instance, Steels (2003) argues that agents' programmed capability of reentering speech production (output) as speech comprehension (input) (i.e., reentrant system) "is useful for detecting and repairing language communication, and thus for pushing language and its underlying meaning toward greater complexity" (p. 11). Similarly, Clowes (2006) argues that inner speech contributes to organizing consciousness. Still, he goes further by also emphasizing the role of inner speech in regulating and shaping ongoing activities and orienting attention. Compared with Steels (2003), Clowes (2006) proposes a self-regulation model in which inner speech could serve "as a scaffold for developing and sustaining cognitive functions beyond the parsing and construction of meaningful and grammatical utterances" (Clowes, 2006, p. 120; see also Clowes, 2007). This model was tested in a series of experiments (Clowes and Morse, 2005) in which groups of agents, implemented with a simple recurrent neural network, had to execute different tasks. The agents in the experimental condition, compared with those in the control condition, were equipped with speech reentrant architectures. Results show that agents with speech reentrance performed better than those who were not programmed with such capability.

Recently, Chella and Pipitone (2020) have proposed a cognitive architecture that uses inner speech to improve robot self-awareness (Chella et al., 2020). It is based on the standard model of mind (Laird et al., 2017) and integrates theoretical contributions on working memory (i.e., phonological loop; Baddeley and Hitch, 1994). The architecture is composed of two main layers: a motor-perception layer and a memory layer. The motor-perception layer enables the robot to perceive information and to take actions: the perception regards data from both itself (e.g., emotions, body, and beliefs and general inner state) and the outside world (the facts in the environment). Along the same line, the motors act on the external entities (e.g., to pick objects, to identify locations) or internal entities (e.g., to self-regulate, to appraise a situation).

The memory layer includes long-term memory, which stores both domain knowledge (declarative memory) and behavioral information (procedural memory), and the working memory system, which models the cognitive functions and processes. The working memory elaborates information from the motor-perception layer by integrating them with information retrieved from the long-term memory.

The architecture of inner speech fits the Baddeley and Hitch's (1994) components into both layers: more specifically, inner speech is modeled as a loop between the phonological store (which briefly holds verbal and written linguistic information) that is a subcomponent of working memory, and the covert articulator (which is responsible to produce linguistic information and then reenter it in the phonological store) located in the motor-perception layer. In Chella and Pipitone's (2020) architecture, inner speech is not just based on a speech reentrance for memorizing data, by which the output word is simply reentered as an input in the phonological store. On the contrary, when the robot processes a word, it is rehearsed by the phonological store, and it is integrated with the information retained in the long-term memory system so that the initial input word is extended in a more elaborated way. All these contributions are aimed at implementing inner speech in automation to improve its functioning (self-awareness, self-regulation, or performance).

No studies, however, have been performed so far with the aim to test if automation equipped with inner speech may affect the quality of HAI and, in particular, human trust toward automation.

THE ROLE OF TRUST IN HAI

Trust research is a well-established field of scientific knowledge that has collected contributions of different disciplines over the years (e.g., psychology, philosophy, sociology; Paliszkiewicz, 2011) and has focused particularly on the field of HAI.

In psychology, trust is a multidimensional concept with no universal definition, which generally refers to an underlying psychological state affected by both cognitive and affective processes (Lewis and Weigert, 1985; McAllister, 1995; Cummings and Bromiley, 1996; Kramer, 1999; Chowdhury, 2005; Johnson and Grayson, 2005; Paliszkiewicz, 2011). Cognitive trust refers to an individual's conscious decision to trust based upon one's beliefs and knowledge about a partner's reliability and competence (McAllister, 1995; Paliszkiewicz, 2011). On the contrary, affective trust stems from interpersonal and emotional bonds, mostly based on the feelings of security, care, and mutual concern (McAllister, 1995; Johnson and Grayson, 2005). From a functional perspective, trust serves as a psychological mechanism for the reduction of social complexity through the formation of expectations and beliefs about others' intentions and behaviors (Luhmann, 1979; Lewis and Weigert, 1985; Rompf, 2014). Lewis and Weigert (1985) state that rational prediction requires time and mental resources for collecting and processing information to determine highly probable outcomes, and thus, trust may be an efficient alternative. Indeed, "by extrapolating past experiences into the future, individuals save the cognitive resources which would be otherwise needed for the search of information and its deliberate processing" (Rompf, 2014, p. 98). Within the psychological literature, trust definitions highlight two key elements: on one side, trust activates positive attitudes, expectations, or confidence in the trustee (Rotter, 1967; Corritore et al., 2003; Lee and See, 2004); on the other, it implies a willingness to put oneself at risk or in a vulnerable state (Mayer et al., 1995; Kramer, 1999; Lee and See, 2004). Muir (1987) states that trust is generally defined as an expectation of or confidence in another and always has a specific referent. Indeed, it involves a relationship between "a trustor A that trusts (judges the trustworthiness of) a trustee B concerning some behavior X in context Y at a time T" (Bauer and Freitag, 2017, p. 2). Moreover, trust is dynamic because it develops and changes over time. Still, it is not a linear process: It may evolve as well as it may deteriorate through a process of loss and repair in response to individual, social, and environmental factors (Paliszkiewicz, 2011; Fulmer and Gelfand, 2013). Similarly, in HAI literature, trust is generally defined as an "attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee and See, 2004, p. 54), and it is considered one of the main factors linked to automation use (Parasuraman and Riley, 1997; Lee and See, 2004; Merritt and Ilgen, 2008; Lewis et al., 2018). Accordingly, trust plays a crucial role in reliance between humans and automation, allowing the latter to take on a full collaborative partner (Lee et al., 2013; Hoff and Bashir, 2015). HAI studies found that people tend to rely on automation they trust and reject those who they do not trust (Muir and Moray, 1996; Lewandowsky et al., 2000; Lee and See, 2004). According to Muir (1987), the same elements that serve as a basis for trust between individuals may affect HAI. However, whereas in human–human interaction, trust is affected by both cognitive and affective processes (Lewis and Weigert, 1985; Mayer et al., 1995), in HAI, trust is considered to be affected predominantly by cognitive aspects because the machine is expected to reach standard performances (Muir, 1994; Merritt and Ilgen, 2008; Lewis et al., 2018).

Lee and Moray (1992) propose that, in HAI, trust is based on three factors: performance, process, and purpose (see also Lee and See, 2004). Performance refers to the automation's capabilities and competencies to achieve the operator's goals. The process focuses on the algorithms and operations that govern the conduct of automation. Purpose concerns the designer's intent behind the automation development. These factors address the user's perception and knowledge of what automation does, how it works, and why it was developed. Merritt and Ilgen (2008) propose a slightly different classification, suggesting four machine characteristics that may affect human trust: competence (i.e., automation's abilities to perform well), responsibility (i.e., automation's functioning information availability to the user), predictability (i.e., automation's behavior consistency), and dependability (i.e., automation's behavior consistency over time). Trust is crucial in HAI because it is related to both misuse and disuse: misuse occurs when humans over-trust automation, relying excessively on its abilities compared with what it can execute, whereas disuse refers to the lack of trust in automation's capabilities so that the human chooses simply not to use it, resulting in a worse outcome (Parasuraman and Riley, 1997; Lee and See, 2004). Both misuse and disuse, especially for high-risk situations, may have catastrophic consequences, such as a plane crash (Lee and See, 2004; Lyons and Stokes, 2012). If trust is appropriately calibrated, which is when human trust correctly matches the true capabilities of the automation (Lee and See, 2004), misuse and disuse may be avoided, enabling an adequate, optimal, and safe HAI (Hoff and Bashir, 2015; Lewis et al., 2018). Nevertheless, there is some evidence that people rely on automation due to a bias that it makes fewer mistakes than humans do, which, in turn, may lead people to reduce reliance on automation as they perceive and remember more automation error and omission than in humans (Dzindolet et al., 2002; Madhavan and Wiegmann, 2004). However, the extent to which an automation error produces changes in the human trust level is still unclear: for instance, trust levels may drop rapidly in response to the first automation errors (Sauer et al., 2016), but they may also decrease when automation fails at humans' easily perceived tasks (i.e., easy error hypothesis; Madhavan et al., 2006) so that the operator infers that, most likely, the automation will not be able to perform a difficult and complex task either. Another similarity in trust among humans and in HAI refers to its development. In this respect, three main phases have been identified: trust formation, dissolution, and restoration (Rousseau et al., 1998; Kim et al., 2009; Fulmer and Gelfand, 2013). Trust formation starts when a trustor chooses to trust a trustee based on the perceived trustworthiness (i.e., ability, benevolence, integrity; Mayer et al., 1995). If trust is repeatedly violated, the trustor decreases trust levels in the trustee, entering the dissolution phase. The restoration phase happens when the trustor deliberately adopts repair strategies that allow trust levels in the trustee to increase again, eventually stabilizing. These three phases are not necessarily linear and straightforward because trust, at some point in time, may be the result of ongoing violations and repairs.

TABLE 1 | Comparison of human–automation and human–robot trust models.

Human–Automation trust factors (Schaefer et al., 2016)	**Human–Robot trust factors** (Hancock et al., 2011)
I–Human-Related	**I–Human-Related**
Cognitive factors (e.g., knowledge and ability)	Ability-Based (e.g., expertise and competency)
Emotive factors (e.g., comfort and confidence)	Characteristics (e.g., personality, age, and gender)
Traits (e.g., personality, age, and gender)	
States (e.g., stress and fatigue)	
II–Automation-Related	**II–Robot-Related**
Features (e.g., intelligence and anthropomorphism)	Attribute-Based (e.g., personality and robot type)
Capability (e.g., behavior, error, and feedback)	Performance-Based (e.g., transparency and failures)
III–Environment-Related	**III–Environment-Related**
Team collaboration (e.g., membership and culture)	Team collaboration (e.g., membership and culture)
Task/Context (e.g., task type and uncertainty)	Tasking (e.g., task type and task complexity)

All these studies highlight that most of the psychological processes involved in human–human interaction can also be accounted for in HAI even though automation may be affected by certain biases because, for instance, they are expected to be infallible. Nevertheless, knowledge of these processes is crucial for improving human trust calibration toward the automation so that their expectancies reflect the actual characteristics and capabilities of the automation, eventually enhancing human–automation collaboration.

MAY AUTOMATION INNER SPEECH AFFECT HUMAN TRUST?

In the HAI literature, most researchers agree that trust dynamically emerges from the exchange of the distinct features of the operator, the machine, and the specific environment in which the interaction takes place (Hancock et al., 2011; Hoff and Bashir, 2015; Schaefer et al., 2016; Kessler et al., 2017; Lewis et al., 2018). Two extensive meta-analyses carried out by Hancock et al. (2011) and Schaefer et al. (2016) identify three main components affecting trust in human–robot interaction and in HAI: human-related factors, automation/robot–related factors, and environment-related factors (see **Table 1**).

In both models, human-related factors take into account individual differences (e.g., personality traits, age, and gender), emotions (e.g., comfort, confidence), and cognition (e.g., expertise, expectancy, and abilities); automation/robot–related factors consist of characteristics such as personality and anthropomorphism and abilities such as behavior, failures, and errors; environment-related factors include elements such as culture, group membership, context, and task.

Our idea is that inner speech could be one of the automation-related factors influencing trust. Inner speech, in particular, could influence the anthropomorphism of automation.

Anthropomorphism is "the act of attributing humanlike qualities to non-human organisms or objects" (DiSalvo and Gemperle, 2003, p. 68), and it incorporates a wide range of human characteristics from poor appearance and basic behaviors to real-like aesthetics and social communication (Pak et al., 2012). In the past years, researchers argue that people might respond socially to computers and other technologies using the same social rules applied to human interaction (Nass et al., 1995; Reeves and Nass, 1996). This phenomenon, named ethopoeia, is defined as an "assignment of human attitudes, intentions, or motives to non-human entities" (Nass et al., 1993, p. 111). Indeed, people might anthropomorphize those robots that behave typically like humans during social interaction (e.g., stare, gestures, etc.; Duffy, 2003). In an experimental study, Salem et al. (2013) confirmed that non-verbal behaviors during social communication affected anthropomorphic inferences about a robot. They found that the robot's coverbal hand and arm gestures during interaction increased participants' anthropomorphic perceptions and likeability, and this effect was greater for the incongruent condition in which the robot's speech and gesture partially matched. Short et al. (2010) find that even the display of a cheating behavior by the robot during a rock–paper–scissors game increased participants' social engagement with the robot and the attributions of mental states.

Robots' appearance and design also have an important influence on their perceived human-likeness: the presence of certain features (i.e., nose, eyelids, and mouth) and width of the head compared with height increases the levels of robot anthropomorphism (DiSalvo et al., 2002). Similarly, Hinds et al. (2004) show that participants delegated responsibility and relied more on the human-like robot coworker (e.g., face, nose, eyes, mouth, and hair) compared with the machine-like robot when performing a task. Several studies show that increasing the anthropomorphism of an interface enhances people's trust in automation aids even when the information presentation and reliability are identical for other non-anthropomorphic interfaces (de Visser et al., 2012; Pak et al., 2012). In addition, van Pinxteren et al. (2019) find that social service robot anthropomorphism (i.e., gaze turn-taking cues) account for significant variation of trust scores. In addition to how an agent looks, people may also respond to how it sounds. Indeed, people tend to trust more those systems that produce human speech rather than synthetic speech (Stedmon et al., 2007; Eyssel et al., 2012).

Similarly, we suggest that automation equipped with an overt mental verbalization system, which reproduces human inner speech, could make it easier for people to attribute human-like qualities to automation, ultimately enhancing human trust in robots.

In our opinion, this might happen for different reasons, which depend both on the effects of monologic/dialogic inner speech in automation's behavior and the overt or covert nature of inner speech.

Considering the first point, we already discussed that some scholars in the field of AI started to implement inner speech in automation to improve their performance, self-regulation, self-awareness, and consciousness (Steels, 2003; Clowes, 2007; Reggia, 2013; Chella et al., 2020).

Thus, a robot equipped with monologic/dialogic inner speech is expected to be more performative, more self-regulated, more aware of its behaviors, and, finally, more similar to humans.

If it is true, as described in the psychological literature, that monologic inner speech influences thinking and reasoning (Vygotsky, 1962; Morin, 2012; Alderson-Day and Fernyhough, 2015; Gade and Paelecke, 2019), automation implemented with inner speech should improve its performance in making decisions that are more complex. For instance, using inner monolog, the automation could self-guide itself in preparing purposes, goals, plans, and test them before acting as people do when self-guiding themselves in tasks that require attention.

Self-guidance and self-instruction could allow automation, similarly to humans, to become more aware of its choices and actions and more self-regulated.

The dialogic inner speech could also influence the robot performance and behavior: if it is true that it helps people to reframe their opinions, taking the others' perspective, and to solve problems, considering them from different points of view, automation equipped with dialogic inner speech should be more able in both perspective taking and problem solving. Besides this, humans could improve their trust in this type of automation.

Considering the second point, we suggest that also the overt or covert nature of inner speech could influence trust. Indeed, covert inner speech is a phenomenon that sometimes is automatic and not visible to others (Martin et al., 2018).

As already described, the process of transforming thinking in voice facilitates problem solving and reading comprehension (Short et al., 1991; Kucan and Beck, 1997). Thus, automation equipped with overt inner speech may perform better than automation equipped with covert inner speech. It could happen because overt inner speech may be reprocessed by automation's vocal recognition systems, sending to the central processor two different inputs: one is the plan of action sent to the output language system, and the second is the plan of action reprocessed by the language input system. These two different inputs may help the automation to have more control and awareness of the sequence of activities planned or executed and, finally, to perform better. Again, the better the automation performance, the higher the human trust.

We also suggest that overt inner speech may help to improve HAI and, in particular, human–automation trust. Indeed, cognitive processes of automation equipped with robot inner speech would be more transparent and more understandable to humans.

For example, during the execution of a cooperative task between the automation and the human, overt inner speech would manifest automation's "mental" processes (e.g., reasoning that underlies its actions, motivation, goals, and plan of actions). In this way, automation becomes a transparent and overt system, letting humans better understand how it works and what determines its behavior.

Mind perception is how people discern between human and nonhuman agents and consists of two core dimensions: (1) agency, e.g., self-control, memory, planning, and communication, and (2) experience, e.g., pain, pleasure, desire, joy, consciousness (Gray et al., 2007). Consequently, transparency in automation cognitive functioning may help people to increase human-like attributions. A recent study shows that transparency about the robot's lack of human psychological processes and abilities reduced children's anthropomorphic tendencies and trust (van Straten et al., 2020). Therefore, transparency may improve automation anthropomorphism, and as described before, automation anthropomorphism influences trust.

Moreover, inner speech makes cooperation more robust because the robot could evaluate different strategies for going out from a stalemate. For example, suppose, for some reason, a step of the whole task to execute is not feasible (e.g., an object to pick is placed in an unattainable position). In that case, the self-dialog may enable the robot to reflect on possible alternatives for reaching the same goal. Meanwhile, the robot can involve the partner in the new planning, and further turns of interaction enrich the cooperation. By hearing the inner reasoning and the evaluation of the robot, the partner may gain more confidence. The robot's behavior becomes not unpredictable, thus affecting the growth of trust.

Overt inner speech could also have a role in the process of trust development. Indeed, humans would probably be more facilitated in developing trust in an overt system, so when dissolution occurs due to errors by automation, a human may better understand why the errors occurred; this, in turn, could facilitate the trust restoration. In this regard, Kim and Hinds (2006) show that, when a robot shows high transparency during an assembling task by explaining its unexpected behaviors, people tend to blame it less compared with those robots who have less transparency.

CONCLUSION

In this paper, we propose the new idea that inner speech could be one of the functions to be implemented in automation to improve its levels of reasoning, thinking, self-control, self-awareness, and, finally, performance. Moreover, we propose that overt inner speech, allowing people that interact with automation to know and understand the reasoning processes that underlie its behaviors, might influence the level of transparency of automation and, finally, its level of anthropomorphism.

Both performance and anthropomorphism are two essential factors influencing human–automation trust, and for these reasons, we consider the implementation of inner speech as a new important frontier for increasing the quality of HAI and the trustworthiness of automation. On the other hand, some

promising cognitive architectures for implementing inner speech have already been proposed. Still, no studies have been performed so far for testing to what extent automation using inner speech affects human trust.

In the end, this paper's discussion is speculative, and it is based exclusively on theoretical considerations based on empirical evidence from different research fields. However, we believe that future studies on inner speech may represent a new frontier in robotics and AI, and in this sense, we hope that our idea may stimulate further research study in this area.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This paper was based upon work supported by the Air Force Office of Scientific Research under award number FA9550-19-1-7025.

REFERENCES

Alderson-Day, B., and Fernyhough, C. (2015). Inner speech: development, cognitive functions, phenomenology, and neurobiology. *Psychol. Bull.* 141, 931–965. doi: 10.1037/bul0000021

Alderson-Day, B., Weis, S., McCarthy-Jones, S., Moseley, P., Smailes, D., and Fernyhough, C. (2016). The brain's conversation with itself: neural substrates of dialogic inner speech. *Soc. Cogn. Affect. Neurosci.* 11, 110–120. doi: 10.1093/scan/nsv094

Baddeley, A. D., and Hitch, G. J. (1974). "Working memory," in *The Psychology of Learning and Motivation: Advances in Research and Theory*, ed G. H. Bower (New York, NY: Academic Press), 47–89.

Baddeley, A. D., and Hitch, G. J. (1994). Developments in the concept of working memory. *Neuropsychology* 8, 485–493. doi: 10.1037/0894-4105.8.4.485

Bauer, P. C., and Freitag, M. (2017). "Measuring trust," in *The Oxford Handbook of Social and Political Trust*, ed E. M. Uslaner (Oxford: Oxford University Press), 1–30.

Bentall, R. P. (1990). The illusion of reality: a review and integration of psychological research on hallucinations. *Psychol. Bull.* 107, 82–95. doi: 10.1037/0033-2909.107.1.82

Chella, A., and Pipitone, A. (2020). A cognitive architecture for inner speech. *Cogn. Syst. Res.* 59, 287–292. doi: 10.1016/j.cogsys.2019.09.010

Chella, A., Pipitone, A., Morin, A., and Racy, F. (2020). Developing self-awareness in robots via inner speech. *Front. Robot. AI* 7:16. doi: 10.3389/frobt.2020.00016

Chowdhury, S. (2005). The role of affect- and cognitions-based trust in complex knowledge sharing. *J. Manag. Issues* 17, 310–326.

Clowes, R. W. (2006). "The problem of inner speech and its relation to the organization of conscious experience: a self-regulation model," in *Proceedings of the AISB06 Symposium on Integrative Approaches to Machine Consciousness*, eds R. Chrisley, R. Clowes, and S. Torrance (Bristol), 117–126.

Clowes, R. W. (2007). A self-regulation model of inner speech and its role in the organization of human conscious experience. *J. Conscious. Stud.* 14, 59–71.

Clowes, R. W., and Morse, A. F. (2005). "Scaffolding cognition with words," in *Proceedings of the Fifth International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems,* eds L. Berthouze, F. Kaplan, H. Kozima, H. Yano, J. Konczak, G. Metta, J. et al. (Lund: Lund University Cognitive Studies), 101–105.

Corritore, C. L., Kracher, B., and Wiedenbeck, S. (2003). On-line trust: concepts, evolving themes, a model. *Int. J. Hum. Comput. Stud.* 59, 737–758. doi: 10.1016/S1071-5819(03)00041-7

Cummings, L. L., and Bromiley, P. (1996). "The Organizational trust inventory (oti): development and validation," in *Trust in Organizations: Frontiers of Theory and Research*, eds R. M. Kramer and T. R. Tyler (Thousand Oaks, CA: Sage), 302–330.

de Visser, E. J., Krueger, F., McKnight, P., Scheid, S., Smith, M., Chalk, S., et al. (2012). The world is not enough: trust in cognitive agents. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* 56, 263–267. doi: 10.1177/1071181312561062

DiSalvo, C., and Gemperle, F. (2003). "From seduction to fulfillment: the use of anthropomorphic form in design," in *Proceedings of the 2003 International Conference on Designing Pleasurable Products and Interfaces* (New York, NY: ACM), 67–72.

DiSalvo, C., Gemperle, F., Forlizzi, J., and Kiesler, S. (2002). "All robots are not created equal: the design and perception of humanoid robot heads," in *Proceedings of the Fourth Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques* (New York, NY: ACM), 321–326.

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Rob. Auton. Syst.* 42, 177–190. doi: 10.1016/S0921-8890(02)00374-3

Dzindolet, M. T., Pierce, L. G., Beck, H. P., and Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Hum. Factors* 44, 79–94. doi: 10.1518/0018720024494856

Eyssel, F., de Ruiter, L., Kuchenbrandt, D., Bobinger, S., and Hegel, F. (2012). "If you sound like me, you must be more human": on the interplay of robot and user features on human-robot acceptance and anthropomorphism," in *Proceedings of the Seventh ACM/IEEE International Conference on Human-Robot Interaction* (New York, NY: ACM), 125–126.

Fernyhough, C. (2004). Alien voices and inner dialog: towards a developmental account of auditory verbal hallucinations. *New Ideas Psychol.* 22, 49–68. doi: 10.1016/j.newideapsych.2004.09.001

Fernyhough, C. (2016). *The Voices Within: The History and Science of How We Talk to Ourselves*. New York, NY: Basic Books.

Frith, C. D., and Done, D. J. (1988). Towards a neuropsychology of schizophrenia. *Br. J. Psychiatry* 154, 437–443.

Fulmer, C. A., and Gelfand, M. J. (2013). "How do I trust thee? dynamic trust patterns and their individual and social contextual determinants," in *Advances in Group Decision and Negotiation: Vol. 6. Models for Intercultural Collaboration and Negotiation*, eds K. Sycara, M. Gelfand, and A. Abbe (Heidelberg: Springer), 97–131.

Gade, M., and Paelecke, M. (2019). Talking matters–evaluative and motivational inner speech use predicts performance in conflict tasks. *Sci. Rep.* 9:9531. doi: 10.1038/s41598-019-45836-2

Gray, H. M., Gray, K., and Wegner, D. M. (2007). Dimensions of mind perception. *Science* 315:619. doi: 10.1126/science.1134475

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., and Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Hum. Factors* 53, 517–527. doi: 10.1177/0018720811417254

Hinds, P. J., Roberts, T. L., and Jones, H. (2004). Whose job is it anyway? a study of human-robot interaction in a collaborative task. *Hum. Comput. Interact.* 19, 151–181. doi: 10.1207/s15327051hci1901and2_7

Hoff, K. A., and Bashir, M. (2015). Trust in automation: integrating empirical evidence on factors that influence trust. *Hum. Factors* 57, 407–434. doi: 10.1177/0018720814547570

Johnson, D., and Grayson, K. (2005). Cognitive and affective trust in service relationships. *J. Bus. Res.* 50, 500–507. doi: 10.1016/S0148-2963(03)00140-1

Kessler, T., Stowers, K., Brill, J. C., and Hancock, P. A. (2017). Comparisons of human-human trust with other forms of human-technology trust. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* 61, 1303–1307. doi: 10.1177/1541931213601808

Kim, P. H., Dirks, K. T., and Cooper, C. D. (2009). The repair of trust: a dynamic bi-lateral perspective and multi-level conceptualization. *Acad. Manag. Rev.* 34, 401–422. doi: 10.5465/amr.2009.40631887

Kim, T., and Hinds, P. (2006). "Who should I blame? effects of autonomy and transparency on attributions in human-robot interactions," in *Proceedings of the 15th International Symposium on Robot and Human Interactive Communication* (Hatfield: IEEE), 80–85.

Kramer, R. M. (1999). Trust and distrust in organizations: emerging perspectives, enduring questions. *Annu. Rev. Psychol.* 50, 569–598. doi: 10.1146/annurev.psych.50.1.569

Kucan, L., and Beck, I. L. (1997). Thinking aloud and reading comprehension research: inquiry, instruction, and social interaction. *Rev. Educ. Res.* 67, 271–299. doi: 10.3102/00346543067003271

Laird, J. E., Lebiere, C., and Rosenbloom, P. S. A. (2017). A standard model of the mind: toward a common computational framework across artificial intelligence, cognitive, science, neuroscience, and robotics. *AI Mag.* 38, 13–26. doi: 10.1609/aimag.v38i4.2744

Larrain, A., and Haye, A. (2012). The discursive nature of inner speech. *Theory Psychol.* 22, 3–22. doi: 10.1177/0959354311423864

Lee, J. D., and Moray, N. (1992). Trust, control strategies, and allocation of function in human-machine systems. *Ergonomics* 35, 1243–1270. doi: 10.1080/00140139208967392

Lee, J. D., and See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Hum. Factors* 46, 50–80. doi: 10.1518/hfes.46.1.50_30392

Lee, J. J., Know, B., Baumann, J., Breazeal, C., and DeSteno, D. (2013). Computationally modeling interpersonal trust. *Front. Psychol.* 4:893. doi: 10.3389/fpsyg.2013.00893

Lewandowsky, S., Mundy, M., and Tan, G. P. A. (2000). The dynamics of trust: comparing humans to automation. *J. Exp. Psychol. Appl.* 6, 104–123. doi: 10.1037//1076-898x.6.2.104

Lewis, J. D., and Weigert, A. (1985). Trust as a social reality. *Soc. Forces* 63, 967–985. doi: 10.2307/2578601

Lewis, M., Scyara, K., and Walker, P. (2018). "The role of trust in human-robot interaction," in *Foundations of Trusted Autonomy*, eds H. A. Abbass, J. Scholz, and D. J. Reid (Cham: Springer), 135–160.

Li, D., Rau, P. P., and Li, Y. (2010). A cross-cultural study: effect of robot appearance and task. *Int. J. Soc. Robot.* 2, 175–186. doi: 10.1007/s12369-010-0056-9

Loevenbruck, H., Grandchamp, R., Rapin, L., Nalborczyk, L., Dohen, M., Perrier, P., et al. (2018). "A cognitive neuroscience view of inner language: to predict and to hear, see, feel," in *Inner speech: New Voices*, eds P. Langland-Hassan and A. Vicente (Oxford: Oxford University Press), 131–167.

Luhmann, N. (1979). *Trust and Power*. New York, NY: Wiley.

Lyons, J. B., and Stokes, C. K. (2012). Human-human reliance in the context of automation. *Hum. Factors* 54, 112–121. doi: 10.1177/0018720811427034

Madhavan, P., and Wiegmann, D. A. (2004). A new look at the dynamics of human-automation trust: is trust in human comparable to trust in machines? *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* 48, 581–585. doi: 10.1177/154193120404800365

Madhavan, P., Wiegmann, D. A., and Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Hum. Factors* 48, 241–256. doi: 10.1518/001872006777724408

Martin, S., Iturrate, I., Millán, J. R., Knight, R. T., and Pasley, B. N. (2018). Decoding inner speech using electrocorticography: progress and challenges toward a speech prosthesis. *Front. Neurosci.* 12:422. doi: 10.3389/fnins.2018.00422

Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. *Acad. Manag. Rev.* 20, 709–734. doi: 10.2307/258792

McAllister, D. J. (1995). Affect- and cognitive-based trust as foundations for interpersonal cooperation in organizations. *Acad. Manag. J.* 38, 24–59. doi: 10.5465/256727

McCarthy-Jones, S., and Fernyhough, C. (2011). The varieties of inner speech: links between quality of inner speech and psychopathological variables in a sample of young adults. *Conscious. Cogn.* 20, 1586–1593. doi: 10.1016/j.concog.2011.08.005

McGuire, P. K., Silbersweig, D. A., Murray, R. M., David, A. S., Frackowiak, R. S., and Frith, C. D. (1996). Functional anatomy of inner speech and auditory verbal imagery. *Psychol. Med.* 26, 29–38. doi: 10.1017/s0033291700033699

Merritt, S. M., and Ilgen, D. R. (2008). Not all trust is created equal: dispositional and history-based trust in human-automation interactions. *Hum. Factors* 50, 194–210. doi: 10.1518/001872008X288574

Morin, A. (2004). A neurocognitive and socioecological model of self-awareness. *Genet. Soc. Gen. Psychol. Monogr.* 130, 197–222. doi: 10.3200/MONO.130.3.197-224

Morin, A. (2011). Self-awareness part 1: definitions, measures, effects, function, and antecedents. *Soc. Personal. Psychol. Compass* 5, 807–823. doi: 10.1111/j.1751-9004.2011.00387.x

Morin, A. (2012). "Inner speech," in *Encyclopedia of Human Behaviors*, ed W. Hirstein (San Diego, CA: Elsevier), 436–443.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *Int. J. Man Mach. Stud.* 27, 527–539. doi: 10.1016/S0020-7373(87)80013-5

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics* 37, 1905–1922. doi: 10.1080/00140139408964957

Muir, B. M., and Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics* 39, 429–460. doi: 10.1080/00140139608964474

Nass, C., Moon, Y., Fogg, B. J., Reeves, B., and Dryer, D. C. (1995). Can computer personalities be human personalities? *Int. J. Hum. Comput. Stud.* 43, 223–239. doi: 10.1006/ijhc.1995.1042

Nass, C., Stewer, J., Tauber, E., and Reeder, H. (1993). "Anthropomorphism, agency, and ethopoeia: Computers as social actors," in *Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems* (New York, NY: ACM), 111–112.

Oleś, P. K., Brinthaupt, T. M., Dier, R., and Polak, D. (2020). Types of inner dialogs and functions of self-talk: comparisons and implications. *Front. Psychol.* 11:227. doi: 10.3389/fpsyg.2020.00227

Pak, R., Fink, N., Price, M., Bass, B., and Sturre, L. (2012). Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics* 55, 1059–1072. doi: 10.1080/00140139.2012.691554

Paliszkiewicz, J. O. (2011). Trust management: literature review. *Management* 6, 315–331.

Parasuraman, R., and Riley, V. (1997). Humans and automation: use, misuse, disuse, abuse. *Hum. Factors* 39, 230–253. doi: 10.1518/001872097778543886

Piaget, J. (1959). *The Language and Thought of the Child*. Hove: Psychology Press.

Reeves, B., and Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places*. Cambridge, MA: Cambridge University Press.

Reggia, J. A. (2013). The rise of machine consciousness: studying consciousness with computational models. *Neural Netw.* 44, 112–131. doi: 10.1016/j.neunet.2013.03.011

Ren, X., Wang, T., and Jarrold, C. (2016). Individual differences in frequency of inner speech: differential relations with cognitive and non-cognitive factors. *Front. Psychol.* 7:1675. doi: 10.3389/fpsyg.2016.01675

Rompf, S. A. (2014). *Trust and Rationality: An Integrative Framework for Trust Research*. New York, NY: Springer.

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *J. Pers.* 35, 651–665. doi: 10.1111/j.1467-6494.1967.tb01454.x

Rousseau, D., Sitkin, S., Burt, R., and Camerer, C. (1998). Not so different after all: a cross-discipline view of trust. *Acad. Manag. Rev.* 23, 393–404. doi: 10.5465/amr.1998.926617

Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., and Joublin, F. (2013). To err is human(-like): effects of robot gesture on perceived anthropomorphism and likeability. *Int. J. Soc. Robot.* 5, 313–323. doi: 10.1007/s12369-013-0196-9

Sauer, J., Chavaillaz, A., and Wastell, D. (2016). Experience of automation failures in training: effects on trust, automation bias, complacency, and performance. *Ergonomics* 59, 767–780. doi: 10.1080/00140139.2015.1094577

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., and Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: implications for understanding autonomy in future systems. *Hum. Factors* 20, 1–24. doi: 10.1177/0018720816634228

Scott, M., Yeung, H. H., Gick, B., and Werker, J. F. (2013). Inner speech captures the perception of external speech. *J. Acoust. Soc. Am.* 133, EL286–EL292. doi: 10.1121/1.4794932

Shergill, S. S., Brammer, M. J., Fukuda, R., Bullmore, E., Amaro, E. Jr., Murray, R. M., et al. (2002). Modulation of activity in temporal cortex during generation of inner speech. *Hum. Brain Mapp.* 16, 219–227. doi: 10.1002/hbm.10046

Short, E., Hart, J., Vu, M., and Scassellati, B. (2010). "No fair!! an interaction with a cheating robot," in *Proceedings of the Fifth ACM/IEEE International Conference on Human-Robot Interaction* (Osaka: IEEE Press), 219–226.

Short, E. J., Evans, S. W., Friebert, S. E., and Schatschneider, C. W. (1991). Thinking aloud during problem solving: facilitation effects. *Learn. Individ. Differ.* 3, 109–122. doi: 10.1016/1041-6080(91)90011-O

Stedmon, A. W., Sharples, S., Littlewood, R., Cox, G., Patel, H., and Wilson, J. R. (2007). Datalink in air traffic management: human factors issues in communications. *Appl. Ergon.* 38, 473–480. doi: 10.1016/j.apergo.2007.01.013

Steels, L. (2003). Language re-entrance and the "inner voice." *J. Conscious. Stud.* 10, 173–185.

Tullett, A. M., and Inzlicht, M. (2010). The voice of self-control: blocking the inner voice increases impulsive responding. *Acta Psychol.* 135, 252–256. doi: 10.1016/j.actpsy.2010.07.008

van Pinxteren, M. M. E., Wetzels, R. W. H., Rüger, J., Pluymaekers, M., and Wetzels, M. (2019). Trust in humanoid robots: Implications for service marketing. *J. Serv. Mark.* 33, 507–518. doi: 10.1108/JSM-01-2018-0045

van Straten, C. L., Peter, J., Kühne, R., and Barco, A. (2020). Transparency about a robot's lack of human psychological capacities: effects on child-robot perception and relationship formation. *ACM Trans. Hum. Robot Interact.* 9, 1–22. doi: 10.1145/3365668

Vygotsky, L. (1962). *Thought and Language*. Cambridge, MA: The MIT Press.

Waters, F., Woodward, T., Allen, P., Aleman, A., and Sommer, I. (2012). Self-recognition deficits in schizophrenia patients with auditory hallucinations: a meta-analysis of the literature. *Schizophr. Bull.* 38, 741–750. doi: 10.1093/schbul/sbq144

Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychol. Rev.* 20, 158–177. doi: 10.1037/h0074428

Whitmer, A. J., and Gotlib, I. H. (2012). Switching and backward inhibition in major depressive disorder: the role of rumination. *J. Abnorm. Psychol.* 121, 570–578. doi: 10.1037/a0027474

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.