HUMAN-ROBOT INTERACTION

Human-like behavioral variability blurs the distinction between a human and a machine in a nonverbal Turing test

F. Ciardo, D. De Tommaso, A. Wykowska*

Variability is a property of biological systems, and in animals (including humans), behavioral variability is characterized by certain features, such as the range of variability and the shape of its distribution. Nevertheless, only a few studies have investigated whether and how variability features contribute to the ascription of humanness to robots in a human-robot interaction setting. Here, we tested whether two aspects of behavioral variability, namely, the standard deviation and the shape of distribution of reaction times, affect the ascription of humanness to robots during a joint action scenario. We designed an interactive task in which pairs of participants performed a joint Simon task with an iCub robot placed by their side. Either iCub could perform the task in a preprogrammed manner, or its button presses could be teleoperated by the other member of the pair, seated in the other room. Under the preprogrammed condition, the iCub pressed buttons with reaction times falling within the range of human variability. However, the distribution of the reaction times did not resemble a human-like shape. Participants were sensitive to humanness, because they correctly detected the human agent above chance level. When the iCub was controlled by the computer program, it passed our variation of a nonverbal Turing test. Together, our results suggest that hints of humanness, such as the range of behavioral variability, might be used by observers to ascribe humanness to a humanoid robot.

INTRODUCTION

In 1950, Alan Turing proposed that, instead of asking the question whether a machine is intelligent or not, we better ask the question of whether a human interrogator ascribes intelligent behavior to the machine or not, especially given that we do not have a good definition of intelligence. In line with this reasoning, Turing proposed a modification of the Imitation Game (*1*) to evaluate whether a machine would be perceived equivalent to a human, in terms of intelligence. The original imitation game is a party game involving three players. Player A is a man, player B is a woman, and player C (who plays the role of the interrogator) is of either sex. In the imitation game, player C is unable to see either player A or player B and can communicate with them only through written notes. By asking questions to player A and player B, player C tries to determine their sex. Turing proposed a version of the game in which player A was replaced by a machine and player C has to determine whether player A and player B are a human or a machine (computer program in the latter case). If the interrogator is unable to determine which answers are given by a human partner and which by a computer program, the latter is said to pass the Turing test—that is, the computer is indistinguishable from a human being. Although the imitation game and the Turing test were originally based on language, they constitute a type of task in which "human judges impartially compare and evaluate outputs from different systems while ignoring the source of the outputs (*2*)." Thus, alternative forms of the imitation game can be based on nonverbal behavior. For instance, Pfeiffer *et al.* (*3*) used a version of a nonverbal Turing test and showed that in gaze-based social interactions with virtual agents, humanness ratings were positively predicted by the frequency with which the virtual agent followed participants' gaze in a contingent manner. Specifically, the authors systematically varied the gaze behavior of an anthropomorphic virtual character along a continuum from a maximal probability of gaze aversion to a maximal probability of gaze following. Participants were asked to judge whether they had been interacting with the human or with the computer program. The results showed that, when the virtual character was introduced as cooperative, humanness ascription was driven by the degree of contingency. Similarly, Willemse *et al.* (*4*) showed that when a robot avatar followed (contingently) participants' gaze, it was judged as more "human-like" than a robot avatar that did not follow the observers' gaze direction.

The nonverbal Turing test has been used in literature (*5*–*8*) as a general concept to denote a test for the ascription of humanness, often departing from the original literal meaning of the Turing test, which initially was designed to test the ascription of intelligence through verbal interaction. More specifically, the general concept of a "Turing test" can be used as a framework to design experimental studies aiming to evaluate what sort of behavioral features should be implemented on an artificial agent to make it possible, or impossible, for a human observer to discriminate a computer program from a human being. When using the term "a variation of a nonverbal Turing test," we refer to this way of understanding the concept.

It is certainly not an easy task for an artificial agent to pass a nonverbal version of the Turing test because the human brain is highly skilled in detecting and discriminating subtle behavioral characteristics of others, such as movement kinematics (*9*, *10*). Natural and artificial agents can be discriminated through biological motion displayed by simple point-light figures (*11*, *12*). Similarly, our brains can recognize emotion and communicative gestures (*13*, *14*) from simple point-light videos.

Given how our brains are tuned to detect very subtle and implicit signals of human-like behavior, it is worth asking whether a humanoid robot can ever pass the nonverbal version of the Turing test by embodying

Social Cognition in Human-Robot Interaction, Italian Institute of Technology, Genoa, Italy.
*Corresponding author. Email: agnieszka.wykowska@iit.it

human characteristics in its physical actions. In the present study, we set out to examine this question by focusing on behavioral variability.

Behavioral variability as a subtle hint of humanness

Behavioral variability is particularly relevant because human behavior, similarly to most biological systems, is indeed highly variable. For example, when repeating several times the same actions both in terms of motor plans and timing, humans' actions will always be characterized by a certain amount of variance. Stergiou and Decker (*15*) showed that in healthy adults, an optimal level of variability is characterized by a particular organization and a chaotic structure. If a biological system deviates from this state, then it will have a low capacity to adapt to perturbations, ending up in either an overly rigid or an unstable system. Wykowska and colleagues (*16*, *17*) tested whether humans are sensitive to human-like behavioral variability of pointing and gaze gestures of a humanoid robot. In two studies, the authors presented to human participants a robot that was programmed to indicate an object on the screen by pointing or gazing at it (*16*, *17*). The authors manipulated the onset time of the gesture to be either completely constant, controlled through a key press of a human (*16*), or preprogrammed but based on prerecorded response times of a human (*17*). Participants were asked to discriminate between the human and the preprogrammed condition. The results showed that participants identified the human conditions (both the human-controlled and prerecorded) above chance, suggesting that the human brain is sensitive to subtle characteristics of human variability in behavior. Participants were not able to indicate which clue they based their judgments on, suggesting that the brain has some sensitivity to human-likeness at lower, implicit levels of processing.

When talking about behavioral variability in humans, it is important to specify that human behaviors are characterized by two critical features: the range of variability and the shape of the distribution. In reaction time (RT) tasks, the range of variability is strictly related to the type of task, its difficulty, and the type of response to be executed, whereas the shape of the distribution across repetitions has a particular feature, namely, it generally fits the ex-Gaussian distribution (*18*, *19*).

Ascription of humanness in interactive tasks

In the context of interaction with others, it has been proposed that behavioral variability might be an evolutionary adaptive mechanism that has guaranteed the success of our species (*20*, *21*). Low variability in actions makes one easily predictable for predators or competitors, reducing chances to succeed or survive in competitive situations. Evidence showed that humans intentionally modulate their behavioral variability to facilitate (or decrease) coordination with a partner. For instance, Vesper *et al.* (*22*) showed that when pairs of participants were asked to synchronize discrete button presses in a RT task (*22*), the SD in their responses was reduced compared with when they produced a simple RT task alone.

Although the abovementioned literature has examined variability in the context of interaction, the topic of the ascription of humanness toward variable behavior in interaction has received little attention so far. More specifically, studies to date have not addressed the question of how behavioral variability affects the ascription of humanness in joint-action settings. According to the second-person neuroscience account (*23*), to understand the mechanisms of social cognition, one needs to use interactive protocols, rather than passive spectatorial paradigms. When passively observing stimuli, the human brain might not activate the same mechanisms as those that are typically activated in daily interactions with others. Following this line of reasoning, ascription of humanness, based on behavioral clues such as variability range and shape of the distribution, might also be affected by whether we actually perform an interactive task with an agent or only observe the agent. In other words, joint action with an artificial agent might play a role in ascription of humanness.

In cognitive neuroscience, joint actions are defined as "any form of social interaction whereby two or more individuals coordinate their actions in space and time to bring about a change in the environment" (*24*). When designing an interactive task with joint action involved, one needs to consider the key cognitive constituents underlying successful joint action: a co-representation of the task and response coordination between partners. Because joint action is grounded in these two mechanisms, they cannot be neglected from analysis when the experimental paradigm involves a joint action task. Therefore, the paper's focus is twofold: the influence of behavioral variability of a robot on ascription of humanness and on the two key cognitive mechanisms underlying joint action, namely, task co-representation and response coordination.

Task (co-)representation

Task (co-)representation is used to refer to the ability to represent one's own actions together with those of others to be able to understand and predict others' behavior. This ability implies that, when performing a task together, co-actors understand what the other's task is and the conditions under which it occurs. Task (co-)representation has been shown to emerge automatically—even when the task would, in principle, allow participants to act without considering and representing the other's task (*25*–*28*). Over the past 15 years, researchers have addressed the mechanism of task (co-)representation with the use of the "joint Simon" paradigm (*27*). The paradigm is a modification of the standard Simon task (*29*–*31*), in which participants are asked to execute key presses with their left and right hand. The respective key presses are assigned to a feature of the target stimulus (a colored square) presented on a screen (example of task instructions: If the square is red, press the left key; if the square is green, press the right key). The target stimuli are presented on the left or right side relative to the center of the screen. This implies that each target stimulus corresponds spatially to one of the responses. In the example above, if the red target is presented on the left, then the response spatially corresponds with the target location, but if the red target is presented on the right, then the response does not correspond with the target location. Performance is typically better (faster and more accurate) in corresponding trials, namely, when the stimulus is presented on the same side of the required response (for example, red square presented on the left), relative to noncorresponding trials, when the stimulus is presented on the opposite side of the required response (for example, red square presented on the right) (*27*, *28*). Such difference in RTs between the noncorresponding and corresponding trials is termed the Simon effect (SE). The SE has not been observed in the go/no-go version of the task—namely, when participants respond only to one feature while withholding the response for the other feature (*29*), for example, when they are asked to respond only to red squares by pressing the left key. However, Sebanz *et al.* (*27*) showed that SE occurs when pairs of participants perform the go/no-go task in a joint context sitting next to each other, with each member of the pair in charge of responding to one color of the target stimulus (the left-seated participant pressing the

left key only when red squares are presented and the right-seated participant pressing the right key only when the green squares are presented). According to the task (co-)representation account (*26*, *27*), the joint SE (JSE) has been interpreted as evidence for shared representation (co-representation): When people perform together complementary parts of a task, they tend to represent the entire task and to integrate both their own and other's action options into a shared representation (see, however, alternative accounts, such as "the reference coding hypothesis" (*32*, *33*), "the task representation hypothesis" (*34*), "the spatial coding theory" (*35*), or "the response coding account" (*36*, *37*).

For the purposes of the present study, it has been shown that the emergence of task (co-)representation can be affected by the type of agent with whom we are interacting in the joint Simon task. For instance, Sahai and colleagues (*38*) tested the emergence of task (co-)representation across human and machine partners. They asked participants to perform the joint Simon task with a confederate or with a computer. Results showed that a JSE was found only when the partner was a human and not when it was a computer [see also (*39*) for similar results and (*40*) for opposite patterns]. Stanzel and colleagues (*41*) replaced one of the two co-agents with the Tombatossals humanoid robot that was described as functioning in either a biologically inspired human-like way or a purely deterministic machine-like manner. Results showed that the JSE emerged in the believed human-like but not in the believed machine-like robot condition. Such a result has been recently extended and generalized by Strait *et al.* (*42*), who designed a joint Simon task across three different countries (Germany, United States, and Mexico) with the NAO robot as a co-agent. The authors reported JSEs with the robot and comparable JSEs across three different countries. Together, this literature review suggests that, within a joint action scenario, humans co-represent actions not only of their conspecifics (other humans) but also of some nonbiological robotic agents. However, the latter is likely to occur only for humanoid robots [experiment 3 in (*43*)].

Response coordination

Another mechanism that allows humans to complete joint action tasks with others is response coordination (*44*). Precise and flexible response coordination is grounded in specific internal models of both self-generated and other-related actions and their integration in real time. Response coordination emerges when partners adapt the timing of their movements to each other. For example, coordination emerges every time we adapt our gait to that of a friend while walking together. This tendency to adapt the timing of our movements to others is called emergent coordination, and it seems necessary to be temporally coupled with others (*44*). Emergent coordination relies on the human capability to extract temporal regularities from external events, use them to predict the subsequent event, and act accordingly (*45*). Thus, the more reliable and consistent the timing of external events, the more reliable and consistent is our prediction about them, and the better we will be able to coordinate. When task co-representations are established, self and other models work together, allowing co-agents to adapt to each other in real time (*45*), thereby eliciting coordination. Malone *et al.* (*46*) compared the response variability structure of participants performing a go/no-go Simon task individually or together with another person. Results showed that the fractal structure of RTs was higher in the individual than in the joint condition (*46*), suggesting that when participants performed the task with a partner, responses were characterized by nested patterns of variability that were not due to random fluctuations (*46*).

Aim of study

The present study aimed to test which aspects of human behavioral variability affect the ascription of humanness to robots during a joint action scenario. Specifically, we were interested to test whether a distribution of behavior that falls within the human-like range (with human-like mean and SD) is a hint that humans use for the ascription of humanness toward a humanoid robot, or whether the robot's behavior must also follow the shape of the distribution that characterizes human behavior (*18*, *19*), namely, the ex-Gaussian distribution. Furthermore, we examined the influence of our manipulation of the robot's behavior on task co-representation and response coordination because these are two inherent aspects of joint action tasks. To address the aims of the study, we designed a joint Simon paradigm in which participants performed the task in pairs; each member of the pair performed the task with a humanoid iCub robot (*47*) along their side. The members of each pair were seated in two separate experimental cabins (Fig. 1 and Movie 1).

The critical manipulation was that the iCub robot could perform the task in either a preprogrammed or a teleoperated manner. In the preprogrammed condition, iCub's behavior fell within the range of human variability but was not human-like in terms of the shape of the distribution. In the teleoperated condition, iCub's button presses were controlled by the other member of the pair in real time, seated in the other room (Fig. 1 and Movie 1). Participants were asked a variation of a Turing test question ("Do you think the robot was controlled by a human?") after every 100 trials.

In each trial, participants were asked to discriminate red or green squares, with each participant being in charge of one color only. For example, the participant seated in the left cabin was required to respond only when a red square was presented regardless of its spatial location, whereas the participant seated in the right cabin was instructed to execute a key press only when a green square was presented, again irrespective of where it was presented. The two robots were responding to the target stimulus complementary to the human partner. Thus, in the above example, the iCub in the left cabin was responding when a green square was presented, whereas the iCub in the right cabin was responding when a red square was presented.

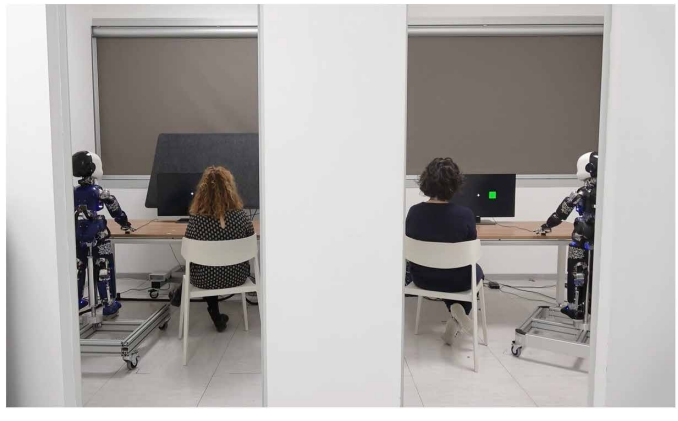


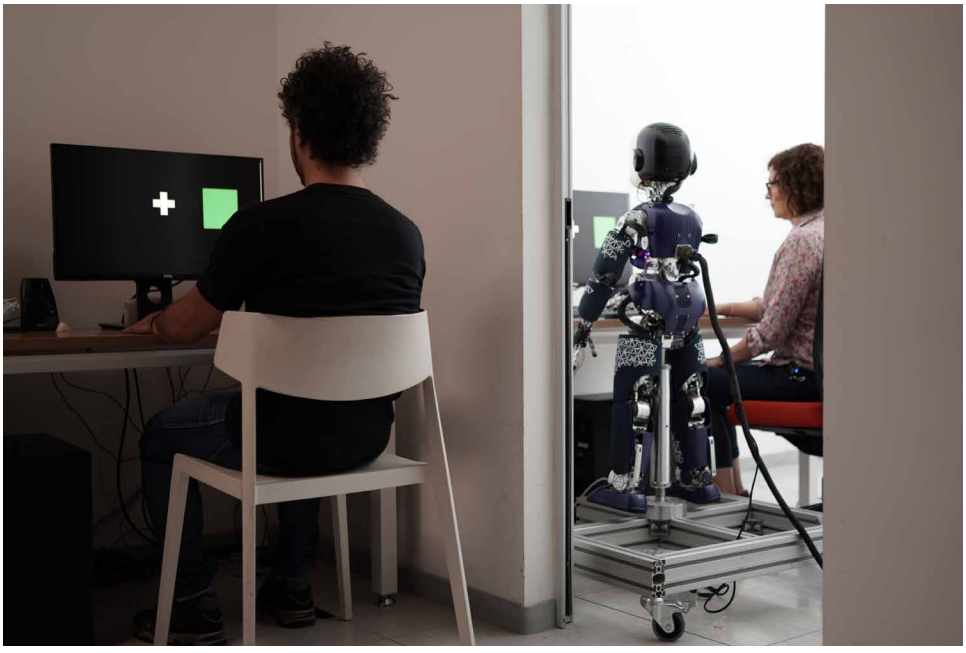
Fig. 1. Experimental setup. Pairs of participants performing the joint Simon task. Each member of a pair performing the task with a humanoid iCub robot along their side.

In the preprogrammed case (computer controller condition), a software generated response times with a uniform (non–human-like shape) distribution based on the mean and SD of human RTs collected during a similar experiment (*48*) (see Materials and Methods for further details). In other words, the behavior of the robot in the computer controller condition was within the human range but with a nonhuman shape of the distribution. In contrast, the human-controlled condition was characterized both by a human-like range of variability and a human-like shape of the distribution (ex-Gaussian) because, in this condition, the robot was in fact controlled by a human (see also Fig. 2).

We reasoned that if the human-like shape of the distribution of RTs is critical for the ascription of humanness, then we should observe a higher proportion of "human" responses in the human-controlled condition and a higher proportion of "computer" responses in the computer-controlled condition, which would indicate that the computer did not pass our variation of a nonverbal Turing test.

On the contrary, if humans use human-like range of variability (but not the shape of the distribution) as a clue for the ascription of humanness, then we should observe in the human condition a higher proportion of "human" responses and under the computer condition either chance-level responses or higher degree of "human" responses, both indicating that the computer passed our variation of a nonverbal Turing test.

Last, to test whether participants were overall sensitive to our manipulation, we evaluated whether distribution of variability differently affected task (co)-representations and emergent coordination because these are implicit mechanisms in joint action.



Movie 1. Overview of the experimental design. This video presents the theoretical background of the study, the concept of the Turing test, and the experimental design in which participants were asked to perform a joint task with a robot and needed to judge whether their robot partner was preprogrammed or controlled by a human.

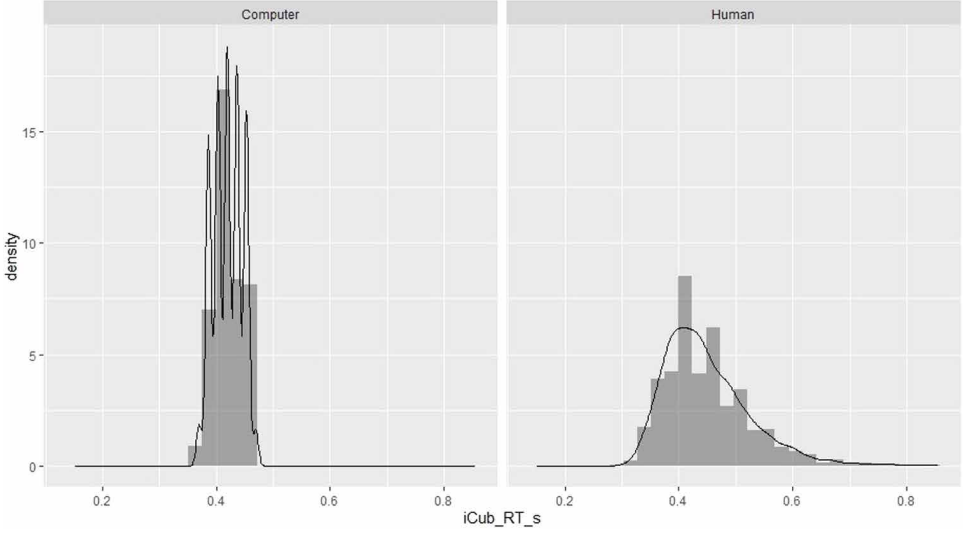


Fig. 2. iCub's RT frequencies during the task. The plots display the RTs of the iCub robot collected during the task as a function of the controller condition (human teleoperated versus preprogrammed computer).

RESULTS

Sensitivity to human variability

To evaluate whether humans are sensitive to human-like behavior during a robot-mediated joint Simon task, the frequencies of correct answers in the Turing question were submitted to a logistic mixed model with controller condition (human versus computer) as a fixed effect and participant as a random effect. Results showed that the probability of responding correctly to the Turing question was higher in the blocks in which the robot was remotely controlled by the human partner compared with when it was run by the computer [$\beta = 1.23$, $z = 3.96$, $P < 0.001$, confidence interval (CI) = (0.63; 1.84)] (Fig. 3). Specifically, the increase in the probability of answering correctly was 3.31 times higher. This shows that participants were sensitive to the subtle humanness clues, namely, the shape of the distribution of RTs in a joint Simon task. To test whether the frequencies of correct answers in the Turing question differed from the chance level (0.5 accuracy rate), we ran one-sample *t* tests on the average accuracy rate for each controller condition against the critical value of 0.5. Results showed that when the robot was controlled by the human agent, participants' accuracy was significantly above the chance level [$t_{23} = 4.30$, $P < 0.001$, $d = 3.36$, CI = (0.60; 0.76)], meaning that

the proportion of human responses was higher than chance. In contrast, under the computer condition, participants accuracy was at the chance level [$t_{23} = -2.04$, $P = 0.053$, $d = 1.40$, CI = (0.27; 0.50)].

Ascription of humanness

To examine further the evidence supporting the idea that the iCub robot passed our variation of a nonverbal Turing test under the computer condition, we submitted the rate of human answers to a one-sample Bayesian *t* test against the critical value of 0.50. Results showed no support for the hypothesis that frequencies of human answers differed from the chance level BF10 = 1.236 (Fig. 4). Thus, in line with our reasoning above, the iCub robot passed our variation of a nonverbal Turing test in the computer controller condition.

Task (co-)representation

To determine whether task (co-)representation emerged in the present task, we first focused on the JSE. Mean RTs for correct trials were modeled as a function of correspondence (NC, noncorresponding; C, corresponding), controller condition (human and computer) and their interactions, as fixed effects, and participants as a random effect. Results showed a main effect of correspondence [$\beta = 10.33$, $t_{(23.90)} = 5.03$, $P < 0.001$, CI = (6.31; 13.36)], indicating faster responses for corresponding than noncorresponding trials (355 ms versus 366 ms, respectively), which indicates a JSE. Also, the main effect of the controller condition was significant [$\beta = 4.42$, $t_{23.90} = 2.17$, $P = 0.030$, CI = (0.42; 8.41)], indicating that participants performed slower when the robot was controlled by the human partner compared with when it was preprogrammed (364 ms versus 358 ms, respectively). The two-way interaction did not reach significance [$\beta = 2.62$, $t < 1$, $P = 0.269$] (Fig. 5), indicating that the JSE was not affected by the controller condition.

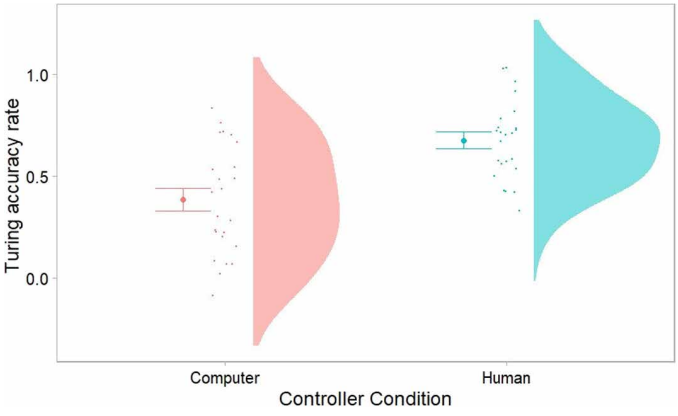


Fig. 3. Logistic mixed model results on the frequencies of correct answers in the Turing question. The plot displays average accuracy rate in the Turing test plotted as a function of controller condition (computer; human) [$\beta = 1.23$, $z = 3.96$, $P < 0.001$, CI = (0.63; 1.84)].

Response coordination

To quantify the degree of coordination with the robot, we estimated instantaneous cross-correlation between each participant and robot. Specifically, we estimated instantaneous cross-correlation (*48*, *49*) between the RT series of participants and the RT series of the iCub with which they were interacting. Subsequently, an index of coupling was estimated as the proportion of correlated activity [the proportion of *r* (correlation coefficient) > 0.25; see (*46*, *48*)] between the RT series of the two agents. The offsets were ±9 trials with a conservative ($\eta = 0.1$) noncausal filter (*49*). Then, the proportion of correlated response activity (index of coupling) was estimated for each participant in each experimental condition and modeled as a function of controller condition (human versus computer) as fixed effects and participant as a random effect. Results showed a lower percentage of instantaneous cross-correlation when the robot was controlled by the human partner compared with when it was running through a computer program (32% versus 36%, respectively) [$\beta = -0.037$, $t_{46} = -5.91$, $P < 0.001$, CI = (−0.05; −0.02)] (Fig. 6).

DISCUSSION

Variability is a property of biological systems. For humans, it is crucial in social interactions. Nevertheless, only a few studies have investigated whether and how it contributes to the ascription of humanness to robots in a human-robot interaction setting. In the present study, we tested whether two aspects of behavioral variability—falling within the human-like range of RT distribution and having

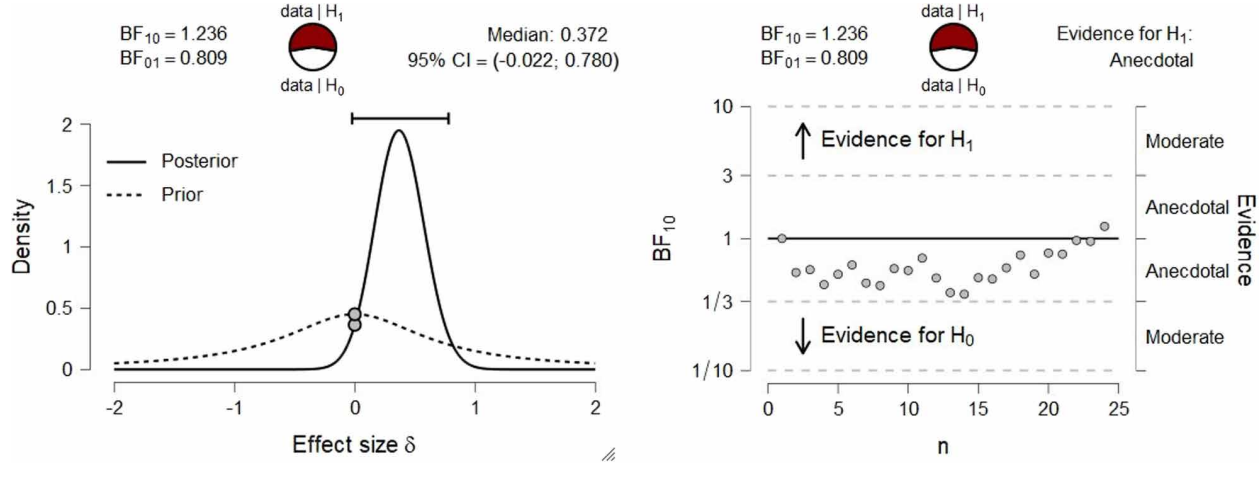


Fig. 4. Bayesian *t* test results. The plots display the posterior and prior distribution of evidence in favor of the null hypothesis (H_0) and the alternative hypothesis (H_1). Bayesian factors (BF) are reported in the figure.

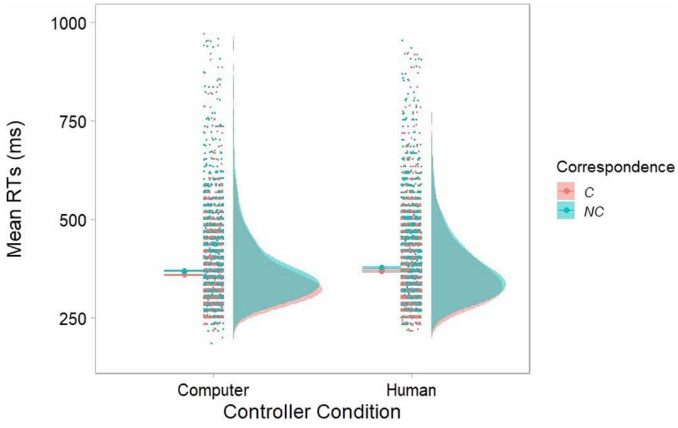


Fig. 5. Task (co-)representation results. The plot displays mean RTs plotted as a function of correspondence (C, corresponding; NC, noncorresponding) across controller condition (computer; human) [$\beta = 10.33$, $t_{23.90} = 5.03$, $P < 0.001$, CI = (6.31; 13.36)].

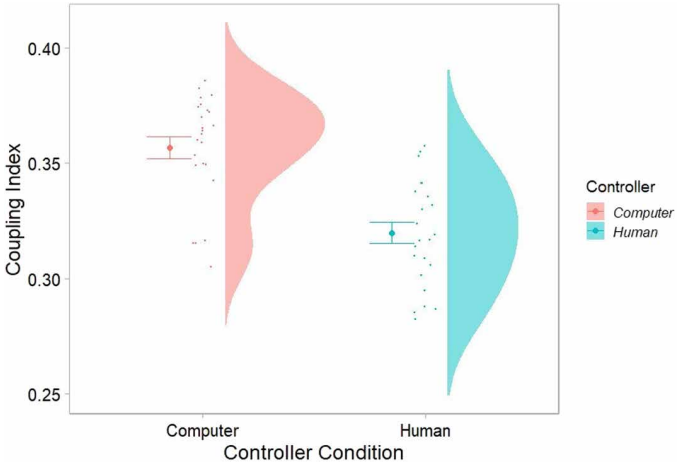


Fig. 6. Response coordination results. The plot displays the average percentage coupling index across controller condition (computer; human) [$\beta = -0.037$, $t_{46} = -5.91$, $P < 0.001$, CI = (−0.05; −0.02)].

the human-like shape of distribution—affect the ascription of humanness to robots during a joint action scenario with a humanoid robot. Also, we examined the influence of variability in the behavior of the robot on two key mechanisms underlying joint action in humans, namely, task co-representation and response coordination.

To address the aims of our study, we designed an interactive task in which pairs of participants were involved in joint action. The members of each pair were seated in two separate experimental cabins. Each member of the pair was performing the joint action task with one iCub robot. The critical manipulation was that either the iCub robot could perform the task in a preprogrammed manner (computer controller condition) or its button presses could be teleoperated in real time by the other member of the pair, seated in the other cabin (human controller condition). In the human controller condition, the RTs had a human-like variability range and human-like shape of distribution (Fig. 2). In the preprogrammed case (computer controller condition), a computer generated response times with a uniform (thus, non–human-like) shape of the distribution but within a human-like range because the preprogrammed behavior was based on human mean and SD (Materials and Methods). When performing the task, participants were asked to answer a variation of a Turing test question, indicating whether they believed that the robot was controlled by a human partner or a computer.

Our results showed that humans are sensitive to human-like response variability, as indicated by the higher accuracy rate in our variation of a nonverbal Turing test for the human controller condition relative to the computer controller condition. When the robot was remotely controlled by the human partner, not only participants responded more correctly to our Turing test question but also their correct response rate was above the chance level.

When the iCub was controlled by the computer program, results showed that participants' responses did not differ from the chance level, indicating that the iCub passed our variation of a nonverbal Turing test because it was difficult for the participants to detect the computer program controlling iCub's behavior. Thus, our results suggest that when interacting with a humanoid robot, behavioral variability that falls within the human-like range but does not have the human-like shape of distribution is used as a clue for the ascription of humanness.

The result that the iCub passed our nonverbal Turing test in the computer controlled condition is also supported by participants' performance during the task. When focusing on participants' RTs during the task, results showed that a statistically significant JSE emerged for both the human and the computer conditions. A significant JSE has been used in cognitive psychology as an index of co-representation of the partner's task in joint action. Namely, even if the action of the partner is not informative or relevant to the individual performance, his/her actions are taken into account. In our study, the JSE was comparable across the two conditions (human versus computer), as indicated by a lack of three-way interaction. This suggests that irrespective of the shape of the distribution of iCub's RTs, participants perceived the task as shared and co-represented iCub's task. Our results corroborate the existing evidence that extends the JSE from human-human to human-robot interaction. Previous studies showed that during a human-robot joint Simon task, beliefs about how the robot is controlled can prevent or promote task (co-) representation in humans (*41*–*43*). For instance, Stenzel and colleagues (*41*) showed that during a human-humanoid joint Simon task, the JSE emerged only when participants were told that the robot was acting in a human-like way but not when they were told that it was acting in a machine-like manner. Thus, in our study, the fact that task co-representation emerged, as evidenced by participants' JSE for iCub also under the computer condition, supports further the notion that iCub passed our variation of a nonverbal Turing test because it was treated like another human would be treated in a joint Simon task (*25*–*29*, *48*).

One could argue that our pattern of results is not due to the robot passing our nonverbal Turing test but is simply driven by the fact that participants' brains were not sensitive to the difference between the two controller conditions. However, results from response coordination showed that the way the iCub was controlled affected emergent coordination differently, with a higher percentage of instantaneous cross-correlation occurring for the computer controller condition compared with the human controller condition. This demonstrates that our manipulation affected participants' performance and that the brain was sensitive to our manipulation.

However, participants did not use this information for their judgment regarding humanness ascription to the iCub. Therefore,

our manipulation affected the brain at the implicit level, without necessarily having an effect on the higher-level cognitive processes that have access to conscious decision-making. This finding suggests that response coordination in joint action is not driven by the conscious ascription of human-likeness but rather by the degree of regularity in behavior. Specifically, participants coordinated better with the iCub when it was controlled by the computer generating RTs with a more uniform distribution compared with when it was controlled remotely by another human, presumably due to a higher degree of predictability of the response times. Although in the computer controller condition, the mean and SD of the distribution were taken from human behavior (*48*), the probability of each RT of the robot was more equally distributed. Thus, the more homogeneous variability of iCub performance under the computer controller might have allowed participants to build a more precise internal model of the timings of its performance, resulting in higher emergent coordination.

Our results also show a dissociation between task (co-)representation, indexed by the JSE, and response coordination. This is an intriguing result. Previous studies that investigated the relationship between the two mechanisms using the joint Simon task suggested that coupling between co-agents may underlie the JSE (*46*, *48*). However, our results showed a dissociation: Response coordination was affected by the dynamics of the behavior of the robot, whereas task co-representation (JSE) was not. This supports the hypothesis that in joint action tasks, co-representation might be a necessary condition for interpersonal coordination but not a sufficient one (*44*).

Overall, our results suggest that, given the current general knowledge of common people regarding robot systems, human-like variability in RTs may be a clue that people use for ascription of humanness to a robot in a joint action scenario such as the joint Simon task. Nevertheless, we cannot exclude that in the near future, when users become more familiar with robots on a daily basis (*50*), behavioral variability might no longer be a feature that plays a role in attributing humanness to artificial agents. For instance, it is possible that by designing robots that are programmed to act with a human-like variability, human observers may start to rely on other cues to distinguish a human-controlled robot from an autonomous one.

Furthermore, it is important to note that in a task in which participants are asked to discriminate between a human and artificial nonverbal behavior, they might not detect humanness per se but rather use features that are shared among all biological systems. Variability is a feature of all biological systems and is not uniquely human. Thus, the nonverbal variation of the Turing test used in our task might have been less specific for measuring the sensitivity to humanness and rather might have tested sensitivity to biological (versus artificial) systems.

Nevertheless, we believe that our results are informative with respect to the design of robot behaviors for human-robot collaborative tasks. For example, for tasks in which it is important that the robot is perceived as human-like, to promote the emergence of task (co-)representation, it is important that its behavior is designed to display human-like variability. Examples of such contexts are work environments in which the human user will be required to monitor or share the workload with the robot. On the other hand, for those contexts in which it is required that the user perceives the robot as a tool and not as a partner, such as during surgery, the robot should be designed to be as predictable (and invariant in behavior) as possible to evoke higher degree of response coordination. Last, at a more theoretical level, our results have implications for the robot version of the Turing test, the total Turing test, as formulated by Harnad (*51*). Harnad objected to the idea that the original Turing test isolates the mind from the body. He proposed that a test for artificial intelligence should also be taking into account the embodiment and the ability to act upon the environment. In this context, our findings suggest that human-like behavioral variability might be a necessary (but not sufficient) condition to pass a nonverbal variation of the total Turing test by an embodied artificial intelligence, which is physically present in the human environment upon which it can also act.

Together, our findings show that the human brain is sensitive to the shape of distribution of the human behavioral variability, perceiving it as a sign of humanness (or of a biological system in general). However, when the behavior of an artificial agent falls within a human-like range (although it has a different shape of distribution), it becomes difficult for a human observer to discriminate between the artificial agent and another human. This provides indications for robot design, which aims at endowing robots with behavior that can be perceived by users as human-like.

MATERIALS AND METHODS

Participants

Thirty-six adults (14 males; 3 left-handed; mean age = 24.1 ± 3.6) were recruited in the experiment. The sample size was defined by an a priori power analysis indicating a sample $n = 24$ to detect a medium effect size [Cohen's d for repeated measures (dz) = 0.41, α (one-tailed) = 0.05 and power = 0.80]. All participants had normal or corrected-to-normal vision and were not informed about the purpose of the study. All participants gave their informed written consent. The studies were conducted under the ethical standards laid down in the 1964 Declaration of Helsinki and were approved by the local ethical committee (Comitato Etico Regione Liguria). Participants received 20€ for their participation. Datasets of five pairs have been excluded from data analysis given the high error rate in the performance of one member of the pair (two pairs) or a technical issue in collecting iCub's response times (three pairs). Therefore, the analysis was run on a sample size of $n = 26$.

Experimental setup and stimuli

The experiment was carried out in two adjacent dimly lit and noiseless rooms. Stimulus presentation, response timing, and data collection were controlled by PsychoPy software (v.2020.1.2). Stimuli were red and green solid squares (2.3° × 2.3°), which were randomly presented on the left or the right of a central white fixation cross (0.6° × 0.6°) on a black background of a computer screen placed on the table in front of each participant (and the corresponding robot partner). Responses were executed by pressing 1.4″ buttons (Logitech Adaptive gaming kit) connected to a MicroPython board working as an HID (human interface device) recognized by the operating system as a standard USB (universal serial bus) keyboard.

Participants were seated facing a 27-inch liquid crystal display screen at a viewing distance of 60 cm. Next to each participant, a full humanoid iCub robot was located. In the left cabin, the robot had a blue torso and arms, whereas in the right cabin, those parts were black. The robots and the controlling PC were connected in the same local area network using Gigabit Ethernet interfaces with different local internet protocol addresses. In this way, we were able to control both robots independently in real time during the execution of

the experiment from the same experimental PC. The action to make the robot press was encoded in the PsychoPy script using the YARP Python wrappers and the IPositionController for controlling the movements of both the iCub robots.

In both conditions, the computer and the human controller, the command to make the robot press was sent at the same occurrence of a specific event, namely, when the visual stimuli of which the robot was in charge to respond appeared. However, they were generated in two different ways.

Under the computer condition, the key press of the robot was generated artificially on the basis of a specific distribution, estimated in a previous study using the joint Simon task in a human-human interactive scenario (*48*). Under this condition, at the beginning of the trial, a value $\boldsymbol{\tau_a}$ was randomly selected from a uniform distribution ($\boldsymbol{\mu}$ = 350 ms, $\boldsymbol{\sigma}$ = 42 ms). Then, when the stimulus was presented on the screen, a command to execute a key press was sent to the robot with a preprogrammed delay of $\boldsymbol{\tau_a}$. Thus, the timestamp of the key press was

$$\text{key press\_ts}_{\text{algorithm}} = \boldsymbol{\tau_a} + \boldsymbol{\delta} \quad (1)$$

with the value $\boldsymbol{\delta}$ representing the latency between the robot and the controlling PC. $\boldsymbol{\delta}$ was measured experimentally, resulting in about 80 ms (Fig. 7).

Under the human condition, the key press of the robot was generated by processing in real time the response of the participant in the other cabin (i.e., the one that was in charge to respond to the same color stimuli of the robot). Whenever the player pressed the button during the assigned trials, the command to execute a key press was sent to the robot in real time. Thus, the timestamp of the key press under the human controller condition was

$$\text{key press\_ts}_{\text{human}} = \boldsymbol{\tau_h} + \boldsymbol{\delta} \quad (2)$$

with the value $\boldsymbol{\tau_h}$ representing the response time of the human participant.

Figure 7 shows the distributions of the latency between the robot and the controlling PC ($\boldsymbol{\delta}$) under both controller conditions, across 200 trials. As presented, in estimating the timestamp of the key press, $\boldsymbol{\delta}$ needed to be taken into account because it represents the delay between when the command for pressing is sent to the robot and when the button is pressed. Of 200 trials, the $\boldsymbol{\delta}$ was pretty much stable for both of the two robots, ranging from 72 to 80 ms. The variability in $\boldsymbol{\delta}$ is due to the refresh rate of the HID input devices used for collecting responses, which was 120 Hz.

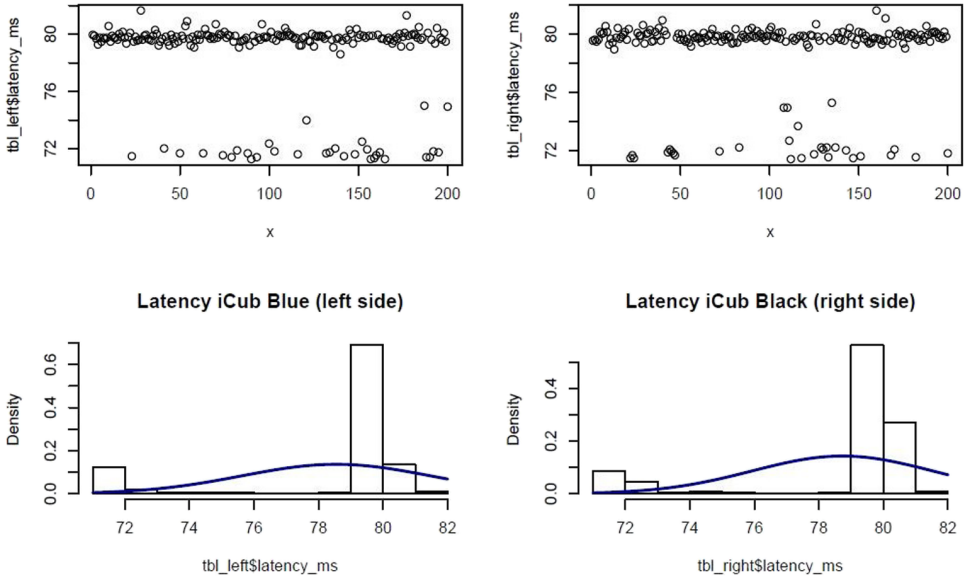


Fig. 7. Latency tests of the two robotic systems. Latencies obtained by testing the two robots systems for a total of 200 trials each. The left column depicts the data related to the robot positioned on the left (iCub blue), and the right column represents the data related to the robot positioned on the right (iCub black). (**Top**) The samples correspond to the latency measures (in milliseconds) for each trial. (**Bottom**) The frequency distributions of the latencies in the same test.

Procedure of the experimental design

Participants were recruited in pairs. The matching of the two members of the pair was made by the experimenter during the recruiting phase. At their arrival at the laboratory, participants filled in three different questionnaires: the instance test (*52*, *53*), the Inclusion of Other in the Self (IOS) scale (*54*), and the Waytz (*55*) intentionality toward robots subscale (Supplementary Materials). Before starting the task, the experimenter presented a demo showing that both robots could be controlled remotely by the person sitting in the other room. The demo was limited to the key press action only. After the demo, participants took their seats in their respective rooms. They were asked to wear earplugs to avoid hearing the sound effect generated by the person sitting in the other room.

A trial began with the presentation of the fixation cross at the center of the screen. After 1 s, the stimulus appeared to the right or the left of the fixation and remained visible until a response was collected or for 800 ms. The maximum time allowed for a response was 1 s after the onset of stimulus presentation. Immediately after a response was collected, or the time of the stimulus presentation elapsed, a black screen was presented for 1 s.

The task consisted of 16 practice trials and 800 experimental trials divided into eight blocks of 100 trials each. For half of the trials, stimulus and response location corresponded (corresponding trials); for the other half, they did not correspond (noncorresponding trials). At the end of each block, participants were asked to answer the question: "Do you think the robot was controlled by a human? Press the button if your answer is yes, do not press if your answer is no." Half of the participants performed first the four human-controlled blocks, followed by the remaining four computer-controlled blocks. For the other half of the participants, the order of blocks was reversed.

Statistical data analysis

Analyses were conducted using the lme4 package (*56*) in R. Parameter estimates (β) and their associated *t* tests (*t* and *P*), calculated using the Satterthwaite approximation for degrees of freedom (*57*), are presented to show the magnitude of the effects, with bootstrapped 95% CIs (*58*).

SUPPLEMENTARY MATERIALS

www.science.org/doi/10.1126/scirobotics.abo1241
Supplementary Methods
Table S1
Fig. S1
MDAR Reproducibility Checklist
Reference (*59*)

REFERENCES AND NOTES

1. A. M. Turing, I.—Computing machinery and intelligence. *Mind* **LIX**, 433–460 (1950).
2. J. H. Moor, The status and future of the Turing test. *Minds Mach.* **11**, 77–93 (2001).
3. U. J. Pfeiffer, B. Timmermans, G. Bente, K. Vogeley, L. Schilbach, A non-verbal turing test: Differentiating mind from machine in gaze-based social interaction. *PLOS ONE* **6**, e27591 (2011).
4. C. Willemse, S. Marchesi, A. Wykowska, Robot faces that follow gaze facilitate attentional engagement and increase their likeability. *Front. Psychol.* **9**, 70 (2018).
5. M. Rebol, C. Güti, K. Pietroszek, Passing a non-verbal Turing test: Evaluating gesture animations generated from speech, in *2021 IEEE Virtual Reality and 3D User Interfaces (VR)* (IEEE, 2021), pp. 573–581.
6. J. Ventrella, M. Seif El-Nasr, B Aghabeigi, R. Overington, Gestural turing test: A motion-capture experiment for exploring believability in artificial nonverbal communication, in *AAMAS 2010 International Workshop on Interacting with ECAs as Virtual Characters* (2010).
7. M. Polceanu. Mirrorbot: Using human-inspired mirroring behavior to pass a turing test, in *2013 IEEE Conference on Computational Inteligence in Games (CIG)* (IEEE, 2013) pp. 1–8.
8. T. Gurion, P. G. Healey, J. Hough, Real-time testing of non-verbal interaction: An experimental method and platform, in *Proceedings of the 22nd Workshop on the Semantics and Pragmatics of Dialogue-Poster Abstracts*, SEMDIAL (2018) pp. 1–4.
9. C. Becchio, L. Sartori, M. Bulgheroni, U. Castiello, Both your intention and mine are reflected in the kinematics of my reach-to-grasp movement. *Cognition* **106**, 894–912 (2008).
10. F. Ciardo, I. Campanini, A. Merlo, S. Rubichi, C. Iani, The role of perspective in discriminating between social and non-social intentions from reach-to-grasp kinematics. *Psych. Res.* **82**, 915–928 (2018).
11. I. M. Thornton, Q. C. Vuong, Incidental processing of biological motion. *Curr. Biol.* **14**, 1084–1089 (2004).
12. E. D. Grossman, R. Blake, Brain areas active during visual perception of biological motion. *Neuron* **35**, 1167–1175 (2002).
13. A. P. Atkinson, W. H. Dittrich, A. J. Gemmell, A. W. V. Young, Emotion perception from dynamic and static body expressions in point-light and full-light displays. *Perception* **33**, 717–746 (2004).
14. T. J. Clarke, M. F. Bradshaw, D. T. Field, S. E. Hampson, D. Rose, The perception of emotion from body movement in point-light displays of interpersonal dialogue. *Perception* **34**, 1171–1180 (2005).
15. N. Stergiou, L. M. Decker, Human movement variability, nonlinear dynamics, and pathology: Is there a connection? *Hum. Mov. Sci.* **30**, 869–888 (2011).
16. A. Wykowska, J. Kajopoulos, M. Obando-Leitón, S. S. Chauhan, J. J. Cabibihan, G. Cheng, Humans are well tuned to detecting agents among non-agents: Examining the sensitivity of human perception to behavioral characteristics of intentional systems. *Int. J. Soc. Robot.* **7**, 767–781 (2015).
17. A. Wykowska, J. Kajopoulos, K. Ramirez-Amaro, G. Cheng, Autistic traits and sensitivity to human-like features of robot behavior. *Interact. Stud.* **16**, 219–248 (2015).
18. R. H. Baayen, P. Milin, Analyzing reaction times. *Int. J. Psych. Res.* **3**, 12–28 (2010).
19. R. M. Yerkes, Variability of reaction-time. *Psychol. Bull.* **1**, 137–146 (1904).
20. D. G. Tervo, M. Proskurin, M. Manakov, M. Kabra, A. Vollmer, K. Branson, A. Y. Karpova, Behavioural variability through stochastic choice and its gating by anterior cingulate cortex. *Cell* **159**, 21–32 (2014).
21. M. D. Fox, A. Z. Snyder, J. L. Vincent, M. E. Raichle, Intrinsic fluctuations within cortical systems account for intertrial variability in human behavior. *Neuron* **56**, 171–184 (2007).
22. C. Vesper, R. P. Van Der Wel, G. Knoblich, N. Sebanz, Making oneself predictable: Reduced temporal variability facilitates joint action coordination. *Exp. Brain Res.* **211**, 517–530 (2011).
23. L. Schilbach, B. Timmermans, V. Reddy, A. Costall, G. Bente, T. Schlicht, K. Vogeley, Toward a second-person neuroscience. *Behav. Brain Sci.* **36**, 393–414 (2013).
24. N. Sebanz, H. Bekkering, G. Knoblich, Joint action: Bodies and minds moving together. *Trends Cogn. Sci.* **10**, 70–76 (2006).
25. S. Atmaca, N. Sebanz, W. Prinz, G. Knoblich, Action co-representation: The joint SNARC effect. *Soc. Neurosci.* **3**, 410–420 (2008).
26. N. Milanese, C. Iani, S. Rubichi, Shared learning shapes human performance: Transfer effects in task sharing. *Cognition* **116**, 15–22 (2010).
27. N. Sebanz, G. Knoblich, W. Prinz, Representing others' actions: Just like one's own? *Cognition* **88**, B11–B21 (2003).
28. N. Sebanz, G. Knoblich, W. Prinz, How two share a task: Corepresenting stimulus-response mappings. *J. Exp. Psychol. Hum. Percept. Perform.* **31**, 1234–1246 (2005).
29. J. R. Simon, A. P. Rudell, Auditory S-R compatibility: The effect of an irrelevant cue on information processing. *J. Appl. Psychol.* **51**, 300–304 (1967).
30. R. W. Proctor, K. P. L. Vu, *Stimulus-Response Compatibility Principles: Data, Theory, and Application* (CRC Press, ed. 1, 2006).
31. M. Tagliabue, M. Zorzi, C. Umiltà, F. Bassignani, The role of long-term-memory and short-term-memory links in the Simon effect. *J. Exp. Psychol. Hum. Percept. Perform.* **26**, 648–670 (2000).
32. T. Dolk, B. Hommel, L. S. Colzato, S. Schütz-Bosbach, W. Prinz, R. Liepelt, How "social" is the social Simon effect? *Front. Psychol.* **2**, 84 (2011).
33. T. Dolk, B. Hommel, W. Prinz, R. Liepelt, The (not so) social Simon effect: A referential coding account. *J. Exp. Psychol. Hum. Percept. Perform.* **39**, 1248–1260 (2013).
34. M. Yamaguchi, H. J. Wall, B. Hommel, Sharing tasks or sharing actions? Evidence from the joint Simon task. *Psychol. Res.* **82**, 385–394 (2018).
35. K. Dittrich, A. Rothe, K. C. Klauer, Increased spatial salience in the social Simon task: A response coding account of spatial compatibility effects. *Atten. Percept. Psychophys.* **74**, 911–929 (2012).
36. C. Iani, F. Ciardo, S. Panajoli, L. Lugli, S. Rubichi, The role of the co-actor's response reachability in the joint Simon effect: Remapping of working space by tool use. *Psychol. Res.* **85**, 521–532 (2021).
37. F. Ciardo, L. Lugli, R. Nicoletti, S. Rubichi, C. Iani, Action-space coding in social contexts. *Sci. Rep.* **6**, 22673 (2016).
38. 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**, 44–55 (2019).
39. T. Wen, S. Hsieh, Neuroimaging of the joint Simon effect with believed biological and non-biological co-actors. *Front. Hum. Neurosci.* **9**, 483 (2015).
40. C. C. Tsai, C. W. J. Kuo, D. L. Hung, O. J. Tzeng, Action co-representation is tuned to other humans. *J. Cogn. Neurosci.* **20**, 2015–2024 (2008).
41. A. Stenzel, E. Chinellato, M. A. T. Bou, A. P. Del Pobil, M. Lappe, R. Liepelt, When humanoid robots become human-like interaction partners: Corepresentation of robotic actions. *J. Exp. Psychol. Hum. Percept. Perform.* **38**, 1073–1077 (2012).
42. M. Strait, F. Lier, J. Bernotat, S. Wachsmuth, F. Eyssel, R. Goldstone, S. Šabanović, A three-site reproduction of the joint Simon effect with the NAO robot, in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (IEEE, 2020) pp. 103–111.
43. A. Sahaï, "Joint agency in human-machine interactions: How to design more cooperative agents?," thesis, PSL Research University (2019).
44. K. L. Marsh, M. J. Richardson, R. C. Schmidt, Social connection through joint action and interpersonal coordination. *Top. Cogn. Sci.* **1**, 320–339 (2009).
45. P. E. Keller, G. Novembre, M. J. Hove, Rhythm in joint action: Psychological and neurophysiological mechanisms for real-time interpersonal coordination. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **369**, 20130394 (2014).
46. M. Malone, R. D. Castillo, H. Kloos, J. G. Holden, M. J. Richardson, Dynamic structure of joint-action stimulus-response activity. *PLOS ONE* **9**, e89032 (2014).
47. G. Metta, L. Natale, F. Nori, G. Sandini, D. Vernon, L. Fadiga, L. Montesano, The *iCub* humanoid robot: An open-systems platform for research in cognitive development. *Neural Netw.* **23**, 1125–1134 (2010).
48. F. Ciardo, A. Wykowska, Response coordination emerges in cooperative but not competitive joint task. *Front. Psychol.* **9**, 1919 (2018).
49. A.V. Barbosa, H. C. Yehia, E. Vatikiotis-Bateson, Algorithm for computing spatiotemporal coordination, in *Auditory and Visual Speech Processing*, S. Luccy, Ed. (Moreton Island: Casual Productions, 2008), pp. 131–136.
50. P. Stone, R. Brooks, E. Brynjolfsson, R. Calo, O. Etzioni, G. Hager, J. Hirschberg, S. Kalyanakrishnan, E. Kamar, S. Kraus, K. Leyton-Brown, D. Parkes, W. Press, A. L. Saxenian, J. Shah, M. Tambe, A. Teller, "Artificial intelligence and life in 2030: One hundred year study on artificial intelligence" (Report of the 2015–2016 study panel, Stanford University, 2016).
51. S. Harnad, Other bodies, other minds: A machine incarnation of an old philosophical problem. *Minds Machines* **1**, 43–54 (1991).
52. S. Marchesi, D. Ghiglino, F. Ciardo, J. Perez-Osorio, E. Baykara, A. Wykowska, Do we adopt the intentional stance toward humanoid robots? *Front. Psychol.* **10**, 450 (2019).
53. F. Ciardo, D. De Tommaso, A. Wykowska, Effects of erring behavior in a human-robot joint musical task on adopting Intentional Stance toward the iCub robot, in *30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)* (IEEE, 2021), pp. 698–703.
54. A. Aron, E. N. Aron, D. Smollan, Inclusion of other in the self scale and the structure of interpersonal closeness. *J. Pers. Soc. Psychol.* **63**, 596–612 (1992).
55. A. Waytz, K. Gray, N. Epley, D. M. Wegner, Causes and consequences of mind perception. *Trends Cogn. Sci.* **14**, 383–388 (2010).
56. D. Bates, R. Kliegl, S. Vasishth, H. Baayen, Parsimonious mixed models. arXiv:1506.04967 [stat.ME] (16 June 2015).

57. A. Kuznetsova, P. B. Brockhoff, R. H. B. Christensen, Package 'lmertest'. *R package version* **2**, 734 (2015).
58. B. Efron, R. J. Tibshirani, *An Introduction to the Bootstrap* (CRC Press, 1994).
59. N. Spatola, A. Wykowska, The personality of anthropomorphism: How the need for cognition and the need for closure define attitudes and anthropomorphic attributions toward robots. *Comput. Hum. Behav.* **122**, 106841 (2021).

Funding: This work has received support from the European Research Council under the European Union's Horizon 2020 research and innovation program, ERC starting grant, G.A. number ERC-2016-StG-715058, awarded to A.W. F.C. was partly supported by H2020 Marie Skłodowska-Curie grant agreement no. 893960. The content of this paper is the sole responsibility of the authors. The European Commission or its services cannot be held responsible for any use that may be made of the information it contains. **Author contributions:** F.C. designed and performed all experiments, analyzed the data, and wrote the manuscript. D.D.T. programmed the robot and revised the manuscript. A.W. designed the experiments and wrote the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data, code, and materials used in the analysis is available at https://osf.io/vyj73/.

Submitted 18 January 2022
Accepted 29 June 2022
Published 27 July 2022
10.1126/scirobotics.abo1241

Science Robotics

Human-like behavioral variability blurs the distinction between a human and a machine in a nonverbal Turing test

F. Ciardo, D. De Tommaso, and A. Wykowska

Sci. Robot. **7** (68), eabo1241. DOI: 10.1126/scirobotics.abo1241

Science Robotics (ISSN 2470-9476) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Robotics* is a registered trademark of AAAS.