International Journal of Social Robotics
https://doi.org/10.1007/s12369-020-00667-4

Perception and Evaluation in Human–Robot Interaction: The Human–Robot Interaction Evaluation Scale (HRIES)—A Multicomponent Approach of Anthropomorphism

Nicolas Spatola[1,2] · Barbara Kühnlenz[3] · Gordon Cheng[3]

Accepted: 10 June 2020

Abstract
The evaluation of how (human) individuals perceive robots is a central issue to better understand human–robot interaction (HRI). On this topic, promising proposals have emerged. However, present tools are not able to assess a sufficient part of the composite psychological dimensions involved in the evaluation of HRI. Indeed, the percentage of variance explained is often under the recommended threshold for a construct to be valid. In this article, we consolidate the lessons learned from three different studies and propose a further developed questionnaire based on a multicomponent approach of anthropomorphism by adding traits from psychosocial theory about the perception of others and the attribution and deprivation of human characteristics: the de-humanization theory. Among these characteristics, the attribution of agency is of main interest in the field of social robotics as it has been argued that robots could be considered as intentional agents. Factor analyses reveal a four sub-dimensions scale including Sociability, Agency, Animacy, and the Disturbance. We discuss the implication(s) of these dimensions on future perception of and attitudes towards robots.

Keywords Robot perception · Robot evaluation · Anthropomorphism · Scale · Questionnaire · Human–robot interaction

1 Introduction

While we are increasingly developing the capabilities of social robots to ensure a broad range of roles they could play in our environment and lives [82, 135], measuring the human perception of those robots and to what extent potential users attribute human characteristics to them is of major importance for the whole design process with regard to the resulting social interaction dynamics between robots and humans [10]. One of those possible social dynamics is the so-called process of anthropomorphism. It describes the attribution of emotional states, competences but also uniquely human traits like morality or rationality to non-humans [33, 127].

As early as 1944, Heider and Simmel investigated the tendency of humans to attribute emotions, motivations, and purpose to simple shapes roaming around in an abstract film in their experiments [53]. Hence, effects of anthropomorphism towards robotic agents that enter the daily lives of humans in order to assist them in manifold complex and cognitive tasks seem to be natural and, thus, have to be considered and evaluated in the design process, dependent on the intended application.

For example, in close human–robot interaction (HRI) in industrial contexts, where humans work in direct physical contact with robots, the same type of motor interferences are observed for incongruent arm movements as in human–human interaction, indicating that observing the actions of humanoid robots rely on similar perceptual processes to observing the actions of human co-workers [94]. The latest research results further suggest, that trying to compensate those motor interferences comes along with increased cognitive task load in comparison to incongruent collaborative arm-movements, conducted by less human-like designed robotic co-workers. Those results hint to the fact that the effects of anthropomorphism, induced by human-like physi-

✉ Nicolas Spatola
NicolasSpatola@hotmail.fr

[1] Social Cognition in Human-Robot Interaction, Istituto Italiano di Tecnologia, Via Morego, 30, 16163 Genova, GE, Italy

[2] Department of Education, University of Potsdam, Am Neuen Palais 10, 14469 Potsdam, Germany

[3] Department of Electrical and Computer Engineering, Institute for Cognitive Systems, Technical University of Munich, Arcisstraße 21, 80333 Munich, Germany

Published online: 13 January 2021

cal robot design may not always be desired and highly depend on the targeted application scenario.

Beyond the design of the physical appearance of a robot, also the extent of human-like behavior, e.g. in trajectory profiles of industrial robots turned out to have a significant positive impact on the health and wellbeing of human users [14]. As an example, minimum-jerk trajectory profiles led to reduced stress levels with regard to heart rate variability in close human–robot collaboration [71], and also the actions of virtual agents have been categorized as more biological by human users when they are animated with motion data captured from human actors in contrast to an interpolation between different poses designed by an animator [12].

In the research field of social robotics, the effects of anthropomorphism are more obvious and can even be used in a targeted way to shape HRI as desired in specific applications, e.g. to induce prosocial behavior towards a robot [72]. Despite the shortcomings of measures for anthropomorphism, the phenomenon itself is well-known and thoroughly investigated with regard to socially interactive robotics [38], and even the extension of legal protections to robotic companions, analogous to animal abuse laws, are discussed [20, 21].

Thus, it is substantial to develop a scientifically valid measure for anthropomorphism in HRI that, in contrast to state-of-the-art measures, considers not only the attribution but also the deprivation of human characteristics, among which agency is of main interest, given that robots can and could be more and more seen as intentional agents in future society [81, 95].

The remainder of the paper is structured as follows: In Sect. 2, the theoretical background is presented with regard to relevant insights from social-cognitive psychology, and state-of-the-art measures of anthropomorphism are discussed. In Sect. 3, the development and validation of the proposed HRIES-questionnaire are presented in a pretest and four consecutive user-studies starting on a morphologic level and different levels of animacy in pictures and videos of state-of-the-art robots representing different levels of human-like design, ending up with real-world HRI in study four. The general limits of the proposed measure are discussed in Sect. 4, and concluding remarks are provided in Sect. 5.

2 Background

2.1 Relevant Insights from Social-Cognitive Psychology

Anthropomorphism is a form of social perception through the attribution of uniquely human traits like morality or rationality to non-humans and goes back to the theory of mentalization. Mentalization is defined as a form of procedural mental activity that energizes the perception and interpretation of the behavior of others in terms of intentional mental states (e.g., beliefs, goals, purposes, and reasons [27]). While the results of recent psychological studies on HRI argue that these different dimensions are mandatory to explain the perception of robots and the attempt to interact with them, actually there is no existing tool for the evaluation of anthropomorphism gathering all those dimensions [35, 70, 112, 113].

Social perception is an evolution construct. To determine whether the other is friends or foe and whether this entity may or not produce a behavior that could help or injure the observer is of prime importance. This theoretical framework has been associated with warmth (e.g., sincerity, trustworthiness, morality) and competence (e.g., ambition, confidence) [29, 40]. The warmth dimension predicts active behaviors such as helping (high warmth) or attacking (low warmth). The competence dimension predicts passive behaviors such as association (high competence) or neglect (low competence). The valence (positive vs. negative) and content (e.g., psychological traits and behaviors) of social evaluation then heavily depend on the degree of perceived warmth and competence associated with the individual or group involved. Individuals or members of social groups stereotyped as warm and competent are perceived much more positively than individuals or members of social groups stereotyped as cold and incompetent. This social evaluation dimensions could also apply, at least in part, to robots [8].

Interestingly, recent research in social robotics proposes that in addition to the basic inter-individual social evaluation dimensions, people could also attribute morality to robots (Banks 2018). This new dimension is not trivial regarding mentalization and the de-humanization theories of Haslam. As we said, mentalization is the extrapolation of a mental reality in others. The inference of mental states to others, including robots, is a particularly important skill for social interactions [13]. The de-humanization taxonomy [50] contains two bi-dimensional constructs. The first one illustrates the attribution of human traits: human uniqueness (e.g., moral sensibility) opposed to animalistic de-humanization (e.g., irrationality), and the second the human nature attribution (e.g., interpersonal warmth) opposed to mechanistic de-humanization (e.g., passivity). According to Haslam's taxonomy of de-humanization, morality and associated traits (e.g. cognitive openness, individuality, depth) are specific human characteristics opposed to a mechanistic conceptualization of the other. The Human nature versus mechanistic de-humanization process refers to the attribution versus deprivation of human characteristics to a fellow creature. In other words, it measures the perceived conceptual distance between the perception of a human and the representation of what is a human. De-humanization is also one of the main psychosocial mechanisms of social acceptance, as a perception of the proximity between others and their group of belonging (in-group) or the self [107]. The observer will use the in-group or him/her-self as the stereotypical representation of the human concept. Studies showed that mechanistic de-humanization results in behavior such as indifference or lack of empathy towards individuals [50, 52, 68]. This de-humanization dimension has proved to be a reliable measure of social evaluation to predict socio-cognitive processes in an HRI situation, especially the Human nature/Mechanistic de-humanization distance measure [112, 114].

2.2 Current Measures of Anthropomorphism

In the HRI-community, two prominent examples of questionnaires developed to measure the attribution of anthropomorphic traits to robots, are widely used: the Godspeed questionnaire series [4], and the Robot Social Attribute Scale (RoSAS) [8].

The Godspeed questionnaires have been a promising first step, however, their development lacks methodology and does not provide any clear test on their structural psychometric validity [58]. As a consequence, the scale is subject to a high variability to its five constructs, i.e. anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety [4]. First, the use of a semantic differential response format (i.e., a bi-dimension scale) relies on a clear identification of the underlying constructs being measured [28]. While some items use antonyms (e.g., dead-alive) others reflect more than a single dimension of judgment (e.g. awful-nice). Thus, the semantic space between the different word pairs cannot be assessed as comparable [67]. Also, there are individual differences in the size and character of the semantic space [55, 56]. Second, several of the opposite items are confounded with positive and negative valence that could explain the high covariance between the dimensions [15]. Finally, the items do not load on factors as proposed in the scale. Factor loadings measure the factorial structure of a scale. They make it possible to group items on separate dimensions and to signify the concept being measured by each factor. On the Godspeed scale, some items load on more than one dimension while others do not load onto any [58]. The result is a significant and extremely high correlation between anthropomorphism, likeability, animacy, and perceived intelligence dimensions (i.e., r = [0.69, 0.89]) suggesting that those concepts have no discriminant validity. They are all measuring the same concept.

Working on these issues, the RoSAS [8] proposes an interesting new dichotomy with the three dimensions: warmth, competence, and disturbance. The authors used a factor analysis on the five dimensions of the Godspeed questionnaires (23 bi-dimensional items, "artificial–lifelike" appearing on both the anthropomorphism and animacy subscale) to reduce it into the three RoSAS' dimensions (18 items). The dimension of warmth and competence are defined as universal dimensions of social perception [40]. These dimensions are central in interpersonal and intergroup perception and are related to cognitive, emotional and behavioral reactions like the tendency to develop empathy or to indulge others. However, regarding the results in real HRI-experiments, the scale produces ambiguous results with a low level of explained variance especially when linked to socio-cognitive processes during HRI [113, 115]. The reason could be that the validation was made using images of robots, which constitutes a different paradigm than actual interaction. Indeed, people do not rely on the same cognitive and neural processes in front of an embodied robot compared to a robot image projected on a screen [66].

In contrast to state-of-the-art approaches, the taxonomy of Haslam as described in Sect. 2.1 was used to evaluate the perception of robots after different types of HRI in recent studies [112, 113]. The authors showed that the attribution of uniquely human traits to robots could be modulated by the form of interaction and could predict the impact of HRI on socio-cognitive processes while the RoSAS couldn't [113]. For example, in one experiment, participants were asked to perform a cognitive control task in the presence of a robot after a social versus non-social HRI. Results showed that in the presence of the social robot, participants performed better on the cognitive control task, an effect called "social facilitation" [112, 114]. In addition, this effect was similar to those observed in the presence of a fellow creature. Results also showed that this effect was moderated by uniquely human versus mechanistic attributions. Indeed, the social facilitation effect was relative to the attribution of human traits on the de-humanization scale. This psychosocial construct was able to provide a deeper perspective than proposed by the dimension of warmth, competence, and disturbance during real HRI. Also, it may help to link the perception of robots to the perception of humanness with fellow creatures. In addition, it could increase the level of variance explained by the RoSAS scale (43.88% according to the main paper). Indeed, the recommended level of explained variance in factor analysis for a construct to be valid is 60% [19, 62, 131].

Thus, in this paper, we propose to improve the RoSAS with new traits from the Haslam's de-humanization theory and to test whether this new scale may precisely measure the perception of robots in HRI psychosocial manipulations. To this end, the following Section presents the development and validation of the proposed HRIES-questionnaire in a pretest and four consecutive user-studies starting on a morphologic level and different levels of animacy in pictures and videos of state-of-the-art robots representing different levels of human-like design, up to real-world interactions with a robot in study four.

3 Development and Validation of the Scale in User Studies

3.1 Pretest

We conducted a pretest to evaluate the semantic redundancy of the different items in order to avoid any weight bias (i.e., redundancy gain effect) of similar items in the scale development [106].

Forty-four items were taken from the Godspeed scale [4], Warmth and Competence dimensions [40], RoSAS (disturbance dimension) [8], De-humanization theory (Human nature dimension) [50] (Table 1) were used in this pretest. To control for potential correlation effect in de-humanization items we conducted a pretest to ensure the independency of positive and negative dimensions. In the pretest, twenty participants (Mage = 20.32, SD 2.03) had to evaluate four robots (i.e., Nao, Yumi, Spot and Meccanoid, see Fig. 1) on the 20 items of Haslam de-humanization taxonomy. Results showed that all pairs were significantly correlated (all $p_s < .05$, $r = [.69, .96]$). Based on these results, we presented only the 5 positive items of the mechanistic de-humanization taxonomy to avoid any semantic differentiator issues [28, 55, 56, 67]. The presentation of semantically dichotomist items tends to energize the emergence a positive/negative judgment effect increasing the likelihood to observe a positive/negative semantic bias in participants responses rather than an in depth treatment of the meaning of the word [58]. Also, the use of positive rather than negative items seems better because of the positivity bias on negative items and the ambiguity that may arise (e.g., higher standard deviation between participants in judgment) [37, 75, 100, 104, 117].

All 44 words were displayed on a computer screen in a randomized table using Qualtrics, an online experiment platform, to 118 English speaking participants recruited on the internet (M_{age} = 25 years, SD 4.13, 79 males, 29 female). They were instructed to “identify if, in the present list, you can find synonyms.” To signify the synonymy of the word, participants had to drag and drop words on their synonyms. We set the threshold of agreement of synonymy to 80%. Groups of items above this threshold were then reduced to one prototypical item.

Seven items were considered as a synonym. To delineate which of the synonym would remain we followed a simple procedure. In each synonymic pair, looking at which word was dragged and which word was used as the referent, we kept the one that was used the most frequently as the referent. All other items were kept for the following study.

Table 1 Synonymy evaluation. The similarity evaluation present, in percentage, the proportion of association of the words among participants

Items	Synonyms	Similarity evaluation
Emotional		
Warm		
Open-mindedness		
Trustworthy	Honest	87%
Friendly	Nice	84%
Likable		
Sincere		
Kind		
Pleasant		
Agency/individuality		
Deep		
Competent		
Intelligent		
Skilled		
Efficient		
Rational	Sensible	83%
Intentional		
Knowlegeable		
Responsible		
Human-like		
Mortal		
Alive		
Real		
Natural	Lifelike	93%
Organic		
Interactive		
Creepy		
Weird	Strange	98%
–	Awkward	87%
Supernatural		
Uncanny	Strange	88%
Freaky		
Shocking		
Eeriness		
Scary		
Dangerous		
Aggressive		
Awful		

3.2 Study 1: Scale Development

The first study aimed to design a structure for the scale. To that end, a fair number of pictures of real robots were evaluated with respect to a combination of de-humanization taxonomy and items of the RoSAS-scale. This picture evaluation method was chosen to provide a holistic approach of robot evaluation in order to extract a first reliable and generalizable matrix. The depiction of robots with different designs allows them to create a higher level of variability

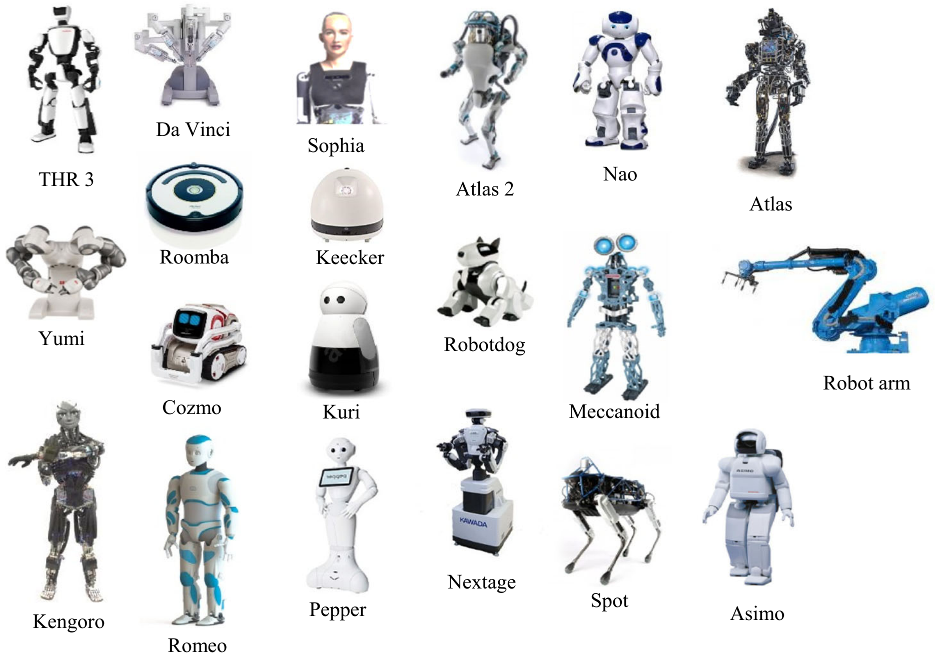

Fig. 1 The 20 robots presented in the questionnaire. Each participant saw a random robot and had to judge 37 traits

and to provide a questionnaire structure that can adapt to a representative sample of robots [96].

3.2.1 Method

The participants were 360 English speakers, recruited on MTurk[1] for 3.00$ (M_{age} = 31 years, SD 8.06, 212 males, 140 female and 8 non-declared). They were informed that they will have to evaluate one of the 20 robots selected for their shape differences on different traits (i.e., "For each trait, you will have to evaluate whether, according to you, it corresponds or not to the robot that is presented to you."). The objective was to create variability in the evaluated stimuli in order to avoid the predominance of loading items on a factor due to the specificity of a robot or type of robot. For each trait a 7-point Likert scale was presented from 1 "not at all" to 7 "totally". The choice of the 7-point Likert scale was motivated by studies about the reliability maximization [39, 97, 98]. Symonds has suggested that reliability is optimized with 7-points scale [120], a suggestion supported by other research (for a review see [18]. The reason would be the limit in the human ability to distinguish between more than seven categories. Lewis also found stronger correlations with *t* test results using 7-point scales [74] considered as an optimum for accurate response [97].

To produce a valid factorial analysis we asked participants to evaluate a random robot out of 20 robots (see Fig. 1) on the 37 selected items (see Table 1). All items were presented in the adjective form in a random order to avoid that participants' responses to questionnaires may be affected by question order [6, 73, 105].

The 20 robots represented a broad range of different design styles and anthropomorphic levels in order to create variability and ensure a generalized use of the scale [131, 133].

3.2.2 Results

Sample Data First, we used Bartlett's sphericity test to ensure inter-item correlation, $\chi^2(666) = 8111.41$, $p<.001$. Inter-item correlations examine the extent to which scores on one item are related to scores on all other items in a scale [17, 129]. Second, we conducted a Kaiser–Meyer–Olkin (KMO) test that verifies that once the linear effect of the other items has been controlled, the partial correlations of each pair of items are low, which would confirm the presence of latent factors linking the items to each other [129]. Its value varies from 0 to 1.1. This is an index for measuring the quality of the data in the sample for the factor analysis. Here the KMO = 0.91. KMO values between 0.8 and 1 indicate the sampling is adequate [9, 30, 60].

Analysis Method In order to determine an initial factorial structure of the scale and sort out unsuitable items, we performed an explanatory factor analysis. We chose a common factor model to attribute the variance to latent factors. This method provides more reliable results than component models (e.g. PCA) in the majority of the cases, while the methods would be roughly equivalent in the remaining cases [25, 46, 109, 126, 128]. Our analysis method started with a

[1] Amazon Mechanical Turk is a crowdsourcing web platform that aims to have humans perform more or less complex tasks for a fee.

principal axis factoring method of extraction with a Promax rotation.[2] The Promax rotation aims to emphasize the differences between the high and low factor saturation coefficients by raising them to the power κ (here 4, the default value[3]). When the loadings are raised to a Kth power, they are all reduced resulting in a simple structure. As the absolute value of the coefficients decreases, the gap between them increases [46, 57, 84]. We conducted analyses on the pattern matrix, which holds the beta weights to reproduce variable scores from factor scores.

Selection of Items The first pattern matrix produced seven factors. We conducted a first exploratory factor analysis (EFA) including all items and used the Kaiser–Guttman–Criterion (eigenvalue > 1) to identify the meaningful number of possible latent factors. For each factor we proceeded as follows: all items were included in a scale reliability analysis to evaluate the reliability of the factor if an item is dropped to maximize the Cronbach's alpha [19, 123]. Negatively correlated items were reversed to control for negative covariance in the Cronbach's alpha equation that incorporates the average of all covariance between items. After the first iteration, we conducted a new iterative EFA with the remaining items until no items could be dropped. From the remaining items, we made a practical choice to maximize the quality of participants' responses saving the reliability of the factors by keeping the four most central items of each factor. Indeed, researchers have demonstrated that length is negatively correlated with the completion and the quality of response of participants [86] especially in self-administered questionnaires [45, 83]. This process made it possible to keep a Cronbach alpha superior to .70 [19, 48] losing the minimum of information. For instance, dropping items loading on the first factor to 4 changed the Cronbach alpha from .94 to .93, optimizing the average variance extracted from .70 to .77. We then conducted a new factorial analysis with the same settings to confirm that the psychometric structure remains the same after each drop of item.

However, it is to mention that such a process reduces the width of the construct to its conceptual centroid. We assume this practical choice to ensure a good balance between practicability and reliability. From the 37 experimental items, 16 remain in the final matrix, $\chi^2(120) = 3237.86, p < .001$; $KMO = .85$, explaining 71.82% of variance (Fig. 2) with 4 factors (Table 2).

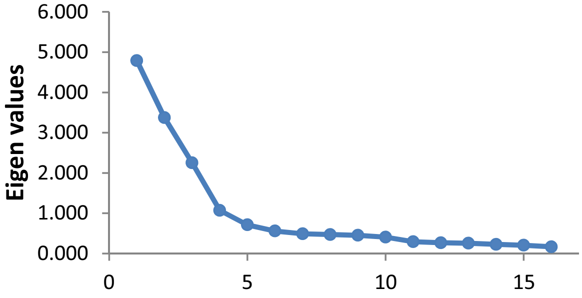

Fig. 2 Eigen values for study 1 factor analysis

3.2.3 Discussion Study 1

The aim of the first study was to delineate a new structure from robot anthropomorphic evaluation based on existing scales proposals and psychosocial theories taxonomy. Our results demonstrate a new taxonomy that may be linked to recent findings in psychology and neuropsychology of HRI. Factorial analysis of the first study showed a structure with four factors.

We found the first factor around "Sociability" attribution including the items "Warm", "Trustworthy", "Friendly", and "Likeable". This construct represents the social constructs that are positively related to the intent of interaction with others [40, 41, 134]. Sociability is the ability of an individual or a group of individuals to evolve in society. Gathering traits perceived as central for human social interactions [65, 76], the Sociability factor, could be determinant for evaluation and acceptance of HRI [22]. Since humans, indeed, tend to evaluate others on their ability to interact positively with them, Fiske and colleagues proposed these attributions as the primary dimension of interpersonal evaluation that they labeled "warmth" [40]. For robots, the conceptualization is relatively different, as should the terminology be. Indeed, this factor is more willing to evaluate the perceived pro-social characteristics rather than intrinsic personality qualities (e.g., moral).

The second factor echoes "Disturbance" attribution including the items "Scary", "Creepy", "Weird", and "Uncanny". This factor represents the negative perception of robots in terms of uncomfortable feelings and perceptions. Unlike the Disturbance dimension in the RoSAS proposal, the Disturbance factor is centered on negative anticipation about something that one cannot consider as "usual". Intrinsically, this dimension represents a form of something negatively unknown resulting in a specific feeling rather than a threat feeling (e.g., "Dangerous" in RoSAS) [78, 79]. For instance, considering robots falling into the uncanny valley is not related to a threat feeling but a feeling of disturbance. Because RoSAS' Disturbance dimension engages two different processes one considering the threat looking or interacting with a robot and one associated with not feeling

[2] Using orthogonal rotation (e.g. VARIMAX), we preserve the independence of the factors. With oblique rotation (e.g. OBLIMIN, PROMAX), we break it and factors are allowed to correlate.

[3] This value is appropriate for most analysis [57].

Table 2 Study 1 pattern matrix presenting loading factors for each item, percent of explained variance and Cronbach's alphas for each factor of the final factors. Items in bold are the items included in the final matrix

Items	Factors			
	1	2	3	4
Warm	**0.910**	0.017	– 0.068	– 0.004
Likeable	**0.866**	– 0.002	– 0.025	0.023
Trustworthy	**0.865**	0.042	– 0.065	0.062
Friendly	**0.863**	– 0.014	– 0.019	0.069
Emotional	0.791	– 0.099	– 0.051	0.038
Pleasant	0.789	0.006	0.176	– 0.127
Kind	0.763	– 0.132	– 0.058	0.046
Open-minded	0.676	0.146	0.078	0.167
Supernatural	– 0.664	0.063	0.157	0.080
Sincere	0.577	0.005	0.298	– 0.120
Eeriness	– 0.515	0.266	0.251	– 0.091
Knowlegeable	0.357	0.227	0.310	0.158
Responsible	0.327	– 0.071	0.275	0.243
Scary	– 0.140	**0.842**	0.067	0.050
Creepy	– 0.113	**0.796**	0.089	0.065
Weird	– 0.168	**0.774**	0.076	0.153
Uncanny	0.161	**0.758**	– 0.093	– 0.060
Awful	– 0.109	0.738	– 0.298	0.053
Shocking	0.161	0.715	– 0.113	– 0.083
Dangerous	0.157	0.628	– 0.043	– 0.093
Freaky	– 0.054	0.621	0.061	– 0.082
Aggressive	– 0.072	0.614	– 0.045	– 0.024
Mortal	0.347	0.415	– 0.164	– 0.087
Rational	– 0.124	– 0.127	**0.903**	0.033
Self-reliant	– 0.099	– 0.031	**0.845**	0.112
Intelligent	0.027	– 0.072	**0.764**	– 0.222
Intentional	0.218	0.187	**0.608**	0.067
Deep	0.003	– 0.455	0.466	0.054
Human-like	0.027	0.113	0.154	**0.647**
Real	– 0.061	– 0.136	0.185	**0.567**
Alive	0.296	– 0.199	– 0.020	**0.504**
Natural	0.279	0.182	0.092	**0.486**
Efficient	– 0.047	– 0.104	– 0.356	0.474
Competent	– 0.035	– 0.191	– 0.307	– 0.045
Skilled	– 0.339	0.325	0.079	0.134
Interactive	– 0.118	0.013	– 0.022	0.015
Organic	0.415	0.048	0.212	– 0.054
% of explained variance	29.940	21.096	14.086	6.695
Cronbach's alpha	0.928	0.879	0.811	0.738

at ease, the evaluation of an uncanny robot would be less efficient because it does not necessary trigger any threat effect. According to our data dangerosity does not seem to be the main predictor. The reason could be that actual robots are not likely to harm people and the feeling of the threat is more distant and abstract (e.g., the fear to be replaced by robots) [1, 119]. Therefore, we assume to not introduce dimension (i.e., threat and disturbance feelings) in a unique factor as, while threat triggers disturbance, disturbance does not necessarily trigger threat.

The third factor is close to the dimension of "Agency" including the items "Rational", "Self-reliant", "Intelligent", and "Intentional". This factor regroups items relevant to the evaluation of the attribution of traits defined as "uniquely human" [50] with a form of agency. The differentiation between the "Competence" dimension [8] and the actual factor echoes previous research on the independence of the two dimensions of capacities to produce a behavior and the mental process behind this behavior. Many studies investigated the perception of robots as intentional agents which is a form of mentalization process [13, 99] that could be conceptualized as a form of co-adaptation [32]. Regarding the social evaluation framework, one component is the evaluation of the others' capacity to act positively or negatively toward the observer. The original items from Fiske and colleagues on the so-called competence dimension were specifically oriented toward high-level cognition traits such as "intelligence" or "determined" rather than "technical" capacities [40]. To treat robots with higher cognitive capacities would be related to consider them as rational agents as proposed by Dennett [26, 27]. This "intentional stance" or "folk psychology stance" is the assumption that an entity, in the present case a robot, will have its own beliefs, thoughts and intents. Therefore, it is a reliable measure of the evaluation of the capacity of a robot to act positively or negatively toward the observer and a reliable measure of the anthropomorphism. As a socio-cognitive process, the more agency, the more social perception and the more anthropomorphic the robot would be seen. Finally, this Agency dimension could be a transitory measure of the "personal stance" which not only relies on the intentional stance but also consider the entity as a person [27, 54, 108].

The fourth factor regroups items associated with "Animacy" shape evaluation: "Human-like", "Real", "Alive", and "Natural", suggesting human characteristics for non-human agents. This factor is close to the general concept of the "Living" oriented to a human form of life. This factor is close to the theoretical concept of "Humanization" as an extension of the simple attribution of human characteristics to a modulation of the conceptual distance between what defines a human for an observer and the attribution to another entity [111].

Taking together Sociability, Agency and Animacy dimensions are positively related to anthropomorphism that is defined in the Oxford dictionary as "the attribution of human traits, emotions, or intentions to non-human entities" [93]. The Disturbance dimension is more ambiguous. Indeed, the disturbance may arise from various factors as demonstrated by the items included in this specific dimension. It goes from a perception of danger, linked to a protective reflex, to a per-

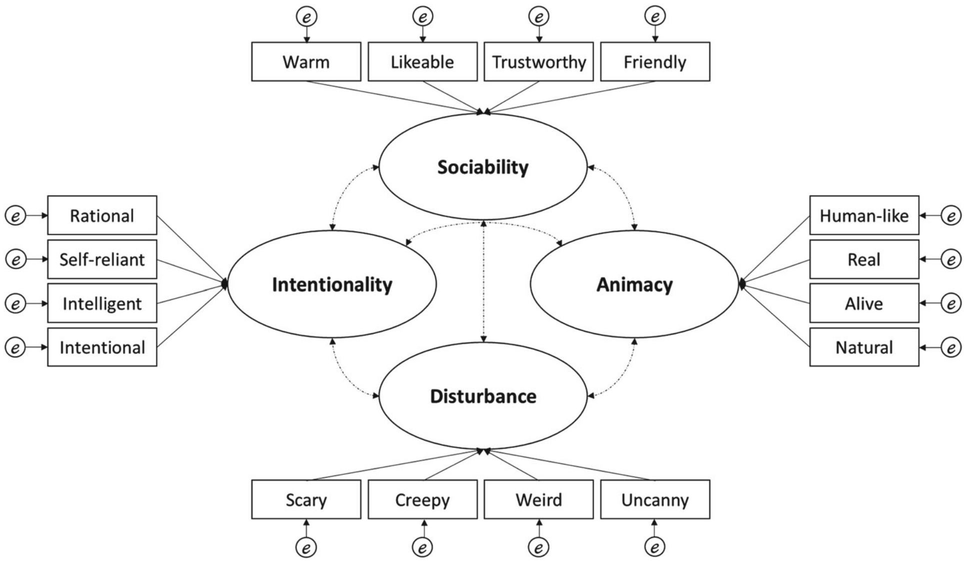

Fig. 3 Measurement theory model (CFA) for the four factors sociability, agency, animacy and disturbance

ception of strangeness, linked to theories such as the uncanny valley [7, 88]. According to the uncanny valley, too anthropomorphic design of a robot or an appearance that does not match with the movements of the robot, but not enough again to blur the difference with a human, it results in a fall-off in acceptance. Thus, Disturbance is more a negative anticipation measure than an anthropomorphic one.

3.3 Study 2: Confirmatory Study

In a second step, we tested the stability of the matrix in the evaluation of a robot in motion. Thus, the second experiment aimed to confirm the study 1 structure of the questionnaire. However, instead of using pictures, participants judge a robot presented on a video. This difference will make it possible to ensure that the scale properties are suitable for both stop and ongoing motion perception. Indeed, the motion may have a great impact on robot perception [11, 64]. For humans and other animals, movement is synonymous with life—so are robots triggering potential positive or negative affects [11, 102].

3.3.1 Method

The participants were 235 English speakers recruited on MTurk[4] for 1.00$ (M_{age} = 20.5 years, SD 6.93, 158 males, 73 female and 4 non-declared). They were informed that they will have to evaluate a robot presented on a short-film on different traits (i.e., "For each trait, you will have to evaluate whether, according to you, it corresponds or not to the robot that is presented to you."). For each trait, a 7-point Likert scale was presented from 1 "not at all" to 7 "totally".

The video presented the NAO robots interacting with a human, an object, and another NAO for 1.36 min. The video came from an Aldebaran Nao presentation video.[5] In order to control from external priming effect, the video was cut to not display any logo and sound. The NAO was chosen for its median human-likeness characteristics norm proposed by the ABOT database (average score = 45.92 on a 100 point scale) [96]. We will develop about this database at continuation.

3.3.2 Results

In order to conduct a confirmatory factor analysis, we checked the Bartlett's sphericity test to ensure inter-item correlation [χ^2 = 1493.61, df = 120, p <.001] and the Kaiser–Meyer–Olkin Indice [KMO = .79] for the sample adequacy [9, 30, 60]. To test the reliability of the proposed structure we conducted a confirmatory factor analysis (CFA) with a structural model using AMOS plugin in SPSS (Fig. 3) using a variance–covariance matrix with maximum likelihood (ML) estimation [87]. ML estimation is more reliable in many cases than others and is widely used [5]. The model-fit indices showed that Chi square (χ^2) value was 189.09 (df = 98, p <0.001). Table 3 shows the recommended model-fit indices [61, 103] as well as the recommended thresholds [130].

As shown in Table 3, all model-fit indices exceeded their respective common acceptance level except for the NFI that was slightly lower than the recommended value. Table 4 presents the non-standardized estimates for each item. All items were significantly associated with their respective factor (all p_s <.001).

[4] Amazon Mechanical Turk is a crowdsourcing web platform that aims to have humans perform more or less complex tasks for a fee.

[5] The original footage can be accessed from https://www.youtube.com/watch?v=rSKRgasUEko.

Table 3 Confirmatory model fit indices. χ^2/df the ratio of Chi square to degree of freedom; GFI the goodness-of-fit-index; AGFI the adjusted goodness-of-fit; NFI the normalized fit index, CFI the comparative fit index; RMSR the root mean square residual

	Recommended value	Values obtained
χ^2/df	$\leq$ 3.00	1.45
GFI	$\geq$ 0.90	0.94
AGFI	$\geq$ 0.80	0.90
NFI	$\geq$ 0.90	0.92
CFI	$\geq$ 0.90	0.97
TLI	$\geq$ 0.90	0.96
RMSEA	$\leq$ 0.08	0.04
SRMR	$\leq$ 0.08	0.07

3.3.3 Discussion Study 2

This second experiment aimed to confirm the structural validity of the new scale. The structural model for the CFA showed a good fit.

3.4 Study 3: Stress Test and Internal Reliability

Recently, Philips and colleagues propose the ABOT (Anthropomorphic roBOT) Database, a collection of real-world anthropomorphic robots [96]. Interestingly, this database proposes a quantification of the human-likeness score for more than 250 robots. We thus used robots (different from the previous ones) from this database based on their human-likeness score. The purpose was (1) to evaluate the psychometric validity and reliability of the new questionnaire using a machine learning approach and (2) to stress the usefulness and reliability of each dimension in the evaluation of the anthropomorphism tendency of participants in regard to social evaluation theories. According to social psychology literature, attitudes are predominantly defined by positive attribution rather than negative attributions [29, 40]. Negative attributions usually occur when there is a lack of positive attributions as neutral/negative attitudes modulators [136]. Therefore, if the present items correctly measure social/anthropomorphic evaluation, the anthropomorphism tendency of participants should be defined first by positive attributions (Agency, Sociability, Animacy), and negative attribution (Disturbance) should act as a modulator when no positive attributions are made. (3) We wanted to evaluate whether the four factors were sensitive to the robot comparison and especially whether the Animacy dimension could follow the human-likeness norms from the ABOT database. (4) Finally, to test external validity, we wanted to put the scale in the perspective of a validated scale: The Negative Attitudes Towards Robots Scale (NARS) [91, 121]. Indeed, positive attitudes towards robots should be positively correlated to positive attribution (Agency, Sociability, Animacy) while negative attitudes should be positively correlated to negative attribution (Disturbance) [33].

3.4.1 Method

The participants were 1086 English speakers recruited by a mailing list (M_{age} = 20.5 years, *SD* 5.71, 246 males, 840 female). They were informed that they will have to evaluate five robots presented on their screen in a random order (i.e., “For each trait, you will have to evaluate whether, according to you, it corresponds or not to the robot that is presented to

Table 4 CFA non-standardized estimates

Items		Factor	Estimate	S.E.	t value	*p* value
Warm	←	Sociability	1.156	0.098	11.758	<.001
Trustworthy	←	Sociability	1.157	0.095	12.133	<.001
Likeable	←	Sociability	1.039	0.098	10.553	<.001
Friendly	←	Sociability	1.287	0.088	14.577	<.001
Scary	←	Disturbance	1.172	0.094	12.491	<.001
Creepy	←	Disturbance	1.215	0.105	11.569	<.001
Uncanny	←	Disturbance	1.349	0.097	13.915	<.001
Weird	←	Disturbance	1.063	0.1	10.621	<.001
Intelligent	←	Agency	1.499	0.09	16.627	<.001
Rational	←	Agency	1.265	0.108	11.725	<.001
Intentional	←	Agency	0.913	0.088	10.355	<.001
Conscious	←	Agency	1.323	0.109	12.107	<.001
Humanlike	←	Animacy	1.008	0.094	10.695	<.001
Alive	←	Animacy	1.144	0.112	10.199	<.001
Natural	←	Animacy	0.908	0.102	8.882	<.001
Real	←	Animacy	1.156	0.098	11.758	<.001

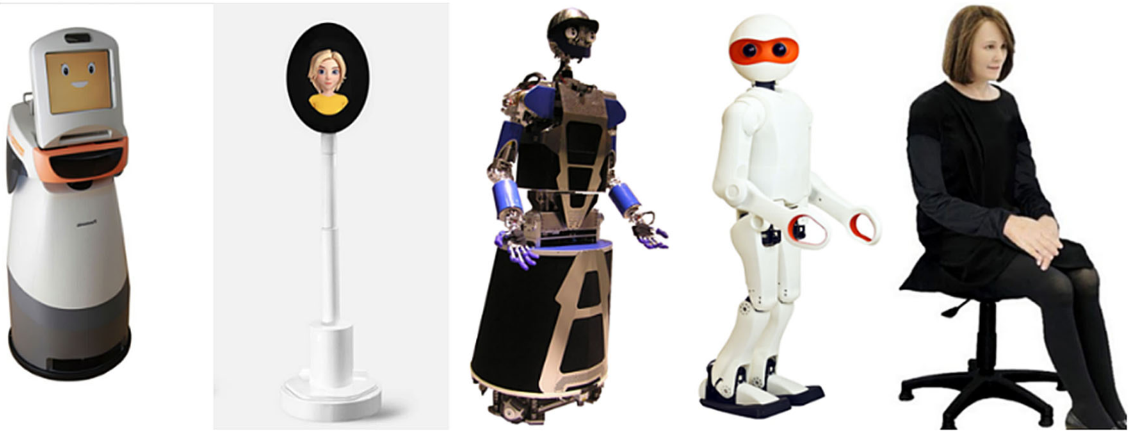

Fig. 4 By order of human-likeness ABOT score, from left to right, Hospi, personal robot, ARMAR, Nimbro, Nadine

you."). For each trait on each robot, a 100-points slider scale was presented from 1 "not at all" to 100 "totally". We chose this 100-points scale to test the reliability of the present factors in the face of more variability in a continuous structure. Some authors have argued that a continuous scale would be better regarding sensitivity, respondent preference [63], and accuracy [24]. Lozano et al. [77] have shown that both the reliability and validity of a Likert Scale decrease when the number of response options is reduced [77]. Slider scale can be used for a greater number of statistical tests and goodness of fit tests may be more powerful compared to a standard Likert scale [44].

The five robots were selected by quintile selection on the human-likeness score of the ABOT database resulting in the use of Hospi, Personal Robot, ARMAR, Nimbro, and Nadine (Fig. 4).

At the end of the experiment, participants completed Nomura, Kanda, Suzuki, and Kato's scale [90] measuring negative attitudes toward robots, hereafter referred to as NARS scale. The NARS scale constitutes of 14 items in three constructs: actual interactions (e.g., "I feel that if I depend on robots too much, something bad might happen") (α = .77); social/future implications (e.g., "I would feel uneasy if robots really had emotions") (α = .63); and emotional attitudes (e.g., "If robots had emotions I would be able to make friends with them") (α = .92). For the purpose of clarity in analysis, we kept the emotional attitudes in its original positive form and did not reverse the scores. For each dimension, participants rated whether they agreed or disagreed (from 1 to 100).

3.4.2 Results

Structural Validity As previously, we checked the Bartlett's sphericity test to ensure inter-item correlation [χ^2 = 57048.83, *df* = 120, *p* <.001] and the Kaiser–Meyer–Olkin Indice (KMO = .75) for the sample adequacy [9, 30, 60]. Again, to test the reliability of the scale we tested a structural model (Fig. 3) using a variance–covariance matrix with maximum likelihood (ML) estimation. The model-fit indices showed that Chi square (χ^2) value was 786.03 (df = 98, *p* <0.001). It is to mention that the χ^2 statistic is very sensitive to sample size [103, 125]. Table 5 shows the recommended model-fit indices [103] as well as the recommended thresholds.

Table 6 presents the non-standardized estimates for each item. All items were significantly associated with their respective factor (all p_s <.001).

To evaluate the structural stability of the present questionnaire we wanted to compare the consistency of the prediction of the scale with a training and test sample. The purpose was to evaluate whether the scale was reliable to reproduce its prediction (here in terms of cluster solution as a respondent profiles proxy) on different samples. We first processed a two-step clustering using Disturbance, Agency, Sociability, and Animacy to delineate anthropomorphic patterns into

Table 5 Confirmatory model fit indices. χ^2/df the ratio of Chi square to degree of freedom; GFI the goodness-of-fit-index; AGFI the adjusted goodness-of-fit; NFI the normalized fit index, CFI the comparative fit index; RMSR the root mean square residual

	Recommended value	Values obtained
χ^2/df	≤ 3.00	4.46
GFI	≥ 0.90	0.96
AGFI	≥ 0.80	0.93
NFI	≥ 0.90	0.97
CFI	≥ 0.90	0.98
TLI	≥ 0.90	0.97
RMSEA	≤ 0.08	0.06
SRMR	≤ 0.08	0.04

Table 6 CFA non-standardized estimates

Items		Factor	Estimate	S.E.	t value	*p* value
Warm	←	Sociability	16.63	0.442	37.641	< .001
Trustworthy	←	Sociability	14.935	0.497	30.067	< .001
Likeable	←	Sociability	18.89	0.442	42.771	< .001
Friendly	←	Sociability	19.457	0.466	41.736	< .001
Scary	←	Disturbance	15.284	1.325	11.534	< .001
Creepy	←	Disturbance	16.171	1.168	13.847	< .001
Uncanny	←	Disturbance	13.832	1.086	12.735	< .001
Weird	←	Disturbance	13.643	1.28	10.654	< .001
Intelligent	←	Agency	14.722	0.601	24.489	< .001
Rational	←	Agency	10.987	0.587	18.712	< .001
Intentional	←	Agency	14.916	0.501	29.752	< .001
Conscious	←	Agency	14.066	0.608	23.137	< .001
Humanlike	←	Animacy	8.869	0.402	22.066	< .001
Alive	←	Animacy	12.088	0.457	26.437	< .001
Natural	←	Animacy	8.234	0.42	19.586	< .001
Real	←	Animacy	8.773	0.957	9.164	< .001

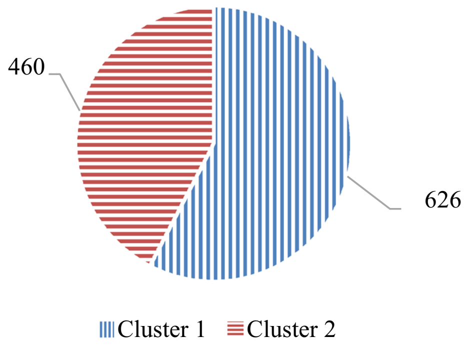

Fig. 5 First solution cluster distribution

participants [2]. The clustering proposed a solution with a 2 clusters' matrice with a 1.36 ratio sizes (Fig. 5) and a cluster quality = 0.5 that measure the cohesion and separation of clusters (good fit).

According to cluster silhouette and cluster comparison, analyses argue for a low versus high anthropomorphism tendency. Indeed, participants in the low cluster attributed less Agency, $F(1, 1085) = 962.45, p < .001, \eta_p^2 = .47$, Sociability, $F(1, 1085) = 1322.67, p < .001, \eta_p^2 = .55$, and Animacy, F(1, 1085) = 881.77, $p < .001$, $\eta_p^2 = .45$, traits to the robots compared to participants in the high anthropomorphism tendency cluster. However, we didn't found difference on Disturbance attribution, $F(1, 1085) = 1.11$, $p = .293$, $\eta_p^2 < .01$ (Table 7).

To evaluate the modulation role of Disturbance, we processed a second cluster analysis including the positive dimensions in a single factor and the Disturbance attribution. We found a 3 cluster solution with a 1.57 ratio sizes (Fig. 6) and a cluster quality = 0.5 (good fit).

Cluster comparison showed that participants in the first cluster presented a higher level of positive attribution compared to both cluster 2 and 3 averaged, $t(1085) = 44.21$, $p < .001$, $d = -2.979$. On that same dimension, the cluster 2 and 3 did not differ, $t(1085) = -1.21, p = .226, d = -0.093$. Interestingly we found a difference between the low clusters in term of negative attribution, $t(1085) = 32.91$, $p < .001$, $d = 2.94$. Also the high level of positive attribution cluster differed from both cluster 2 and 3 averaged, $t(1085) = -.51$, $p = .611$, $d = 0.03$ (Table 8).

Second, we used a machine learning approach to evaluate the cluster predictive reliability of the questionnaires'

Table 7 First cluster solution. Centroids in function of cluster and factors. Factors are presented by order of importance for the clustering solution from left to right

	Agency		Social		Human-Likeness		Disturbance	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Cluster								
1	23.08	12.16	17.21	11.37	15.48	8.38	30.81	17.94
2	44.88	10.39	44.03	12.83	32.44	10.43	29.73	14.86
Combined	32.31	15.72	28.57	17.89	22.66	12.52	30.35	16.71

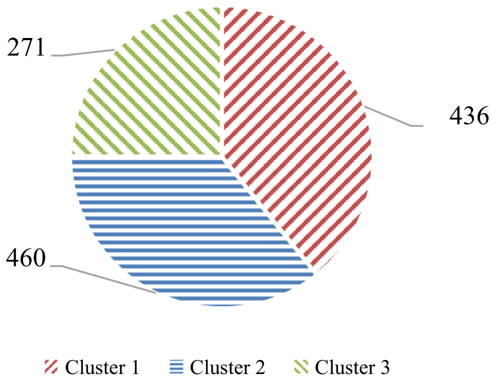

Fig. 6 Second solution cluster distribution

Table 8 Second cluster solution

	Positive attributions		Disturbance	
	Mean	SD	Mean	SD
Cluster				
1	41.01	7.46	31.79	13.63
2	19.66	8.46	16.81	8.22
3	18.91	7.56	47.52	12.99
Combined	27.85	13.18	30.35	16.71

Centroids in function of cluster and factors. Factors are presented by order of importance for the clustering solution from left to right

items on the first cluster solution, it is to say, the reliability of the questionnaire to predict whether an individual will tend to anthropomorphize or not. Data were divided into a 0.80 split. We trained the model on 810 participants and test it on 216. The training phase aims to delineate the predictive value of factors (items) in regard to the high/low anthropomorphism tendency cluster solution. The test subset is used to evaluate whether the actual model reliably predicts the cluster appurtenance of participants. The algorithm predicts the appurtenance of the test participants and compares the prediction to the actual cluster appurtenance of the test participants.

We first trained the predictive model using a Multivariate adaptive regression splines (MARS) algorithm [43]. The model reached 97.22% accuracy to predict high versus low anthropomorphism cluster appurtenance of test subjects using the present questionnaire (Table 9).

Robot Comparison For each anthropomorphic dimension, we conducted a repeated measure ANOVA including the five robots as within factor. Results for each dimension are presented at continuation.

Disturbance Our scale was sensitive enough to discriminate between the 5 different robots in terms of Disturbance traits attribution, $F(4{,}1138) = 246.00$, $p < .001$, $\eta_p^2 = .18$.

Table 9 The confusion matrix is a matrix that measures the quality of a classification system

	Confusion matrix	
	Estimated class	
	High	Low
Real class		
High	132	3
Low	3	78
Fit indices		
Sensitivity		0.998
Specificity		0.963
Pos Pred value		0.978
Neg Pred value		0.963
Precision		0.998
Recall		0.978
F1		0.978
Prevalence		0.625
Detection rate		0.611
Detection prevalence		0.625
Balanced accuracy		0.970
Kappa		0.941

Each line corresponds to a real class, each column corresponds to an estimated class. The fit indices present the characteristics of the predictive solution

Table 10 Study 3 contrasts on disturbance dimension in function of robots

	Personal	Armar	Nimbro	Nadine
Hospi	$F(1, 1141) = 236.41$, $p < .001$, $\eta_p^2 = .17$	$F(1, 1141) = 1036.40$, $p < .001$, $\eta_p^2 = .48$	$F(1, 1141) = 9.17$, $p = .015$, $\eta_p^2 = .01$	$F(1, 1141) = 501.46$, $p < .001$, $\eta_p^2 = .31$
Personnal	x	$F(1, 1141) = 320.15$, $p < .001$, $\eta_p^2 = .22$	$F(1, 1141) = 264.72$, $p < .001$, $\eta_p^2 = .19$	$F(1, 1141) = 153.99$, $p < .001$, $\eta_p^2 = .12$
Armar	x	x	$F(1, 1141) = 1266.59$, $p < .001$, $\eta_p^2 = .53$	$F(1, 1141) = 6.35$, $p = .119$, $\eta_p^2 = .01$
Nimbro	x	x	x	$F(1, 1141) = 600.58$, $p < .001$, $\eta_p^2 = .35$

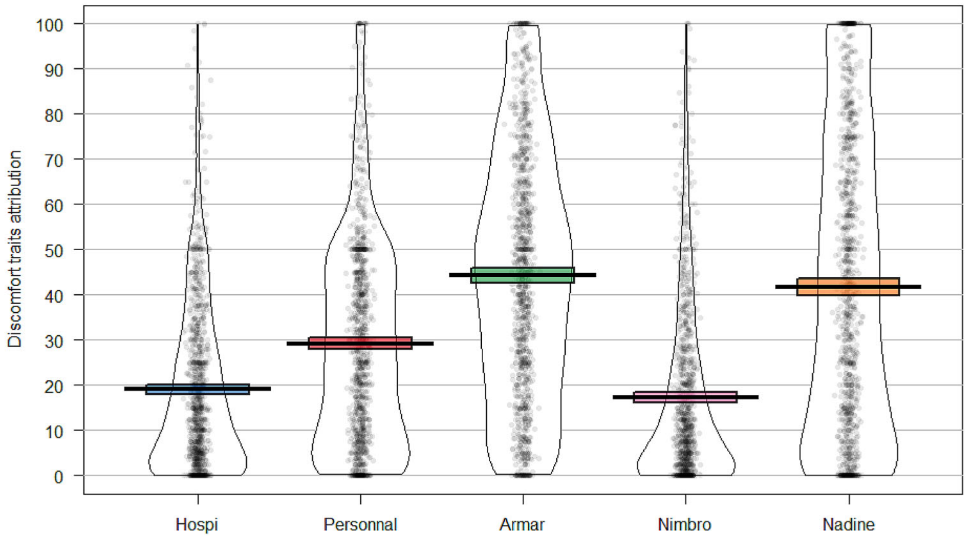

Fig. 7 Disturbance traits average score and distribution in function of the type of robot

Table 11 Study 3 contrasts on agency dimension in function of robots

	Personal	Armar	Nimbro	Nadine
Hospi	$F(1, 1141) = 24.08$, $p < .001$, $\eta_p^2 = .02$	$F(1, 1141) = 91.19$, $p < .001$, $\eta_p^2 = .07$	$F(1, 1141) = 371.85$, $p < .001$, $\eta_p^2 = .25$	$F(1, 1141) = 676.05$, $p < .001$, $\eta_p^2 = .37$
Personnal	x	$F(1, 1141) = 163.68$, $p < .001$, $\eta_p^2 = .13$	$F(1, 1141) = 465.38$, $p < .001$, $\eta_p^2 = .29$	$F(1, 1141) = 802.48$, $p < .001$, $\eta_p^2 = .41$
Armar	x	x	$F(1, 1141) = 94.44$, $p < .001$, $\eta_p^2 = .08$	$F(1, 1141) = 392.82$, $p < .001$, $\eta_p^2 = .26$
Nimbro	x	x	x	$F(1, 1141) = 154.45$, $p < .001$, $\eta_p^2 = .12$

Contrasts are presented in Table 10 with Bonferroni correction (Fig. 7).

Agency The robot were also accurately discriminated on Agency traits attribution, $F(4{,}1138) = 261.84$, $p < .001$, $\eta_p^2 = .48$. Contrasts are presented in Table 11 with Bonferroni correction (Fig. 8).

Sociability According to the scale, the robots were also different in term of Sociability traits attribution, $F(4{,}1138) = 251.47$, $p < .001$, $\eta_p^2 = .47$. Contrasts are presented in Table 12 with Bonferroni correction (Fig. 9).

Animacy Finally, the robots were different in term of Animacy traits attribution, $F(4{,}1138) = 736.34$, $p < .001$, $\eta_p^2 = .72$, following the ABOT database pattern according to the present scale. Contrasts are presented in Table 13 with Bonferroni correction (Fig. 10).

Construct Validity To test the external validity of the present scale we compared the level of anthropomorphic attribution to the attitudes towards robots of participants. We expected a strong correlation between the NARS and the present questionnaire as attitudes toward robots should predict, in part, anthropomorphic attribution. We processed *Pearson correlation* analyses including NARS dimensions, Disturbance, Agency, Sociability, and Animacy factors. The results are presented in Table 14.

3.4.3 Discussion

The present study aimed to validate the psychometric reliability of the new questionnaire and evaluate the sensitivity of the 4 dimensions on 5 new robots from a validated database.

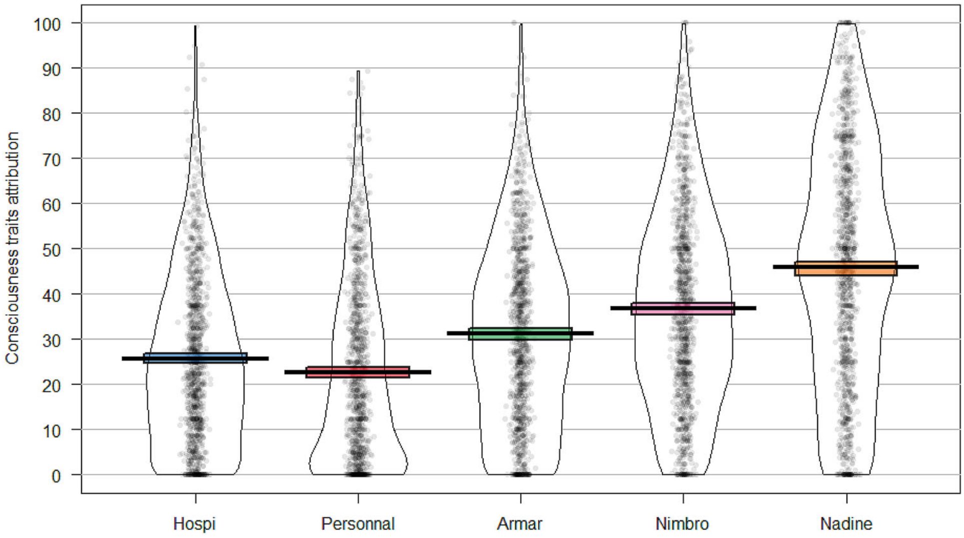

Fig. 8 Agency traits average score and distribution in function of the type of robot

Table 12 Study 3 contrasts on sociability dimension in function of robots

	Personal	Armar	Nimbro	Nadine
Hospi	$F(1, 1141) = 288.70$, $p < .001$, $\eta_p^2 = .20$	$F(1, 1141) = 806.07$, $p < .001$, $\eta_p^2 = .41$	$F(1, 1141) = 20.38$, $p < .001$, $\eta_p^2 = .02$	$F(1, 1141) = 182.22$, $p < .001$, $\eta_p^2 = .14$
Personnal	x	$F(1, 1141) = 152.29$, $p < .001$, $\eta_p^2 = .12$	$F(1, 1141) = 162.95$, $p < .001$, $\eta_p^2 = .13$	$F(1, 1141) = 1.179$, $p = .278$, $\eta_p^2 < .01$
Armar	x	x	$F(1, 1141) = 676.73$, $p < .001$, $\eta_p^2 = .37$	$F(1, 1141) = 165.36$, $p < .001$, $\eta_p^2 = .13$
Nimbro	x	x	x	$F(1, 1141) = 106.5$, $p < .001$, $\eta_p^2 = .09$

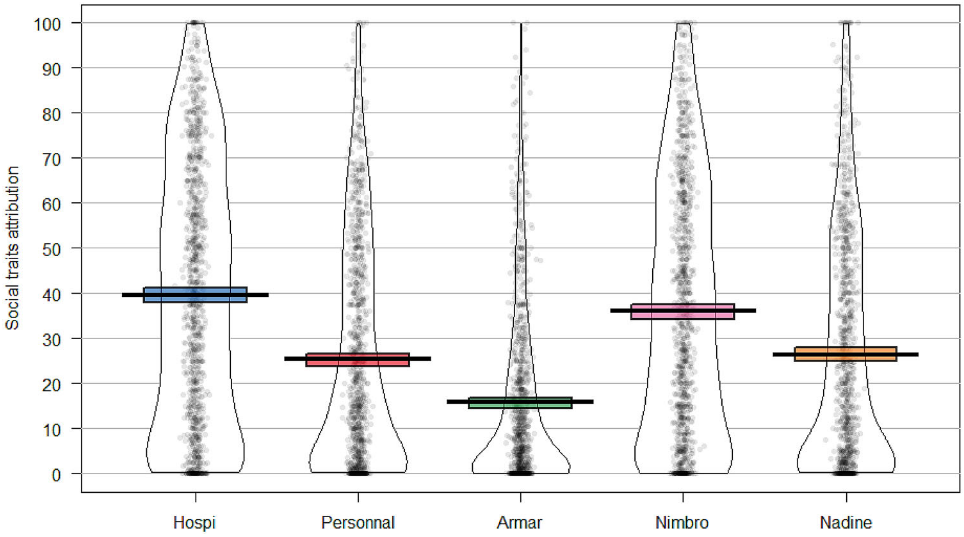

Fig. 9 Social traits average score and distribution in function of the type of robot

Factorial analysis again validated the 4 dimensions structure of the questionnaire. The cluster split argues for a basic dichotomic perception were positive traits (Agency, Sociability, Animacy) rely on a common dimension and negative traits (Disturbance) act as a modulator in a second step. The machine learning approach makes it possible to test the reliability and stability of the questionnaire reaching a 97.22% prediction to classify high versus low anthropomorphism tendency group appurtenance of participants. This result confirms the underlying common anthropomorphic dimension for positive attribution. Finally, we used a continuous scale to evaluate the reliability of the questionnaire without categorical responses.

The use of the continuous scale does not seem to change the structure or the stability of the constructs. According to Cicchetti and colleagues eight, nine, ten, or even 100-point

Table 13 Study 3 contrasts on Animacy dimension in function of robots

	Personal	Armar	Nimbro	Nadine
Hospi	$F(1, 1141) = 16.66, p < .001, \eta_p^2 = .01$	$F(1, 1141) = 104.27, p < .001, \eta_p^2 = .08$	$F(1, 1141) = 471.93, p < .001, \eta_p^2 = .29$	$F(1, 1141) = 2719.28, p < .001, \eta_p^2 = .71$
Personnal	x	$F(1, 1141) = 18.73, p < .001, \eta_p^2 = .02$	$F(1, 1141) = 188.63, p < .001, \eta_p^2 = .14$	$F(1, 1141) = 2497.55, p < .001, \eta_p^2 = .69$
Armar	x	x	$F(1, 1141) = 168.59, p < .001, \eta_p^2 = .13$	$F(1, 1141) = 2333.47, p < .001, \eta_p^2 = .67$
Nimbro	x	x	x	$F(1, 1141) = 1750.19, p < .001, \eta_p^2 = .61$

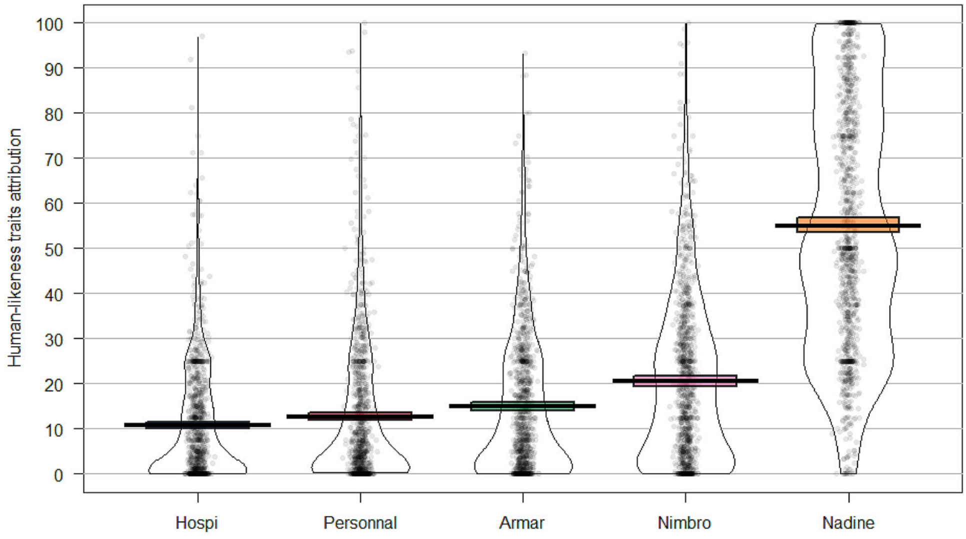

Fig. 10 Human-like traits average score and distribution in function of the type of robot

Table 14 Correlation between negative attitude towards robots scale and disturbance, agency, sociability and animacy dimension

	r	*t*	*p*
Disturbance			
Social/future implications	0.321	11.148	0.000
Emotional attitudes	0.306	10.569	0.000
Actual interactions	− 0.126	− 4.174	0.000
Agency			
Social/future implications	− 0.044	− 1.461	0.144
Emotional attitudes	− 0.037	− 1.230	0.219
Actual interactions	0.170	5.676	0.000
Sociability			
Social/future implications	− 0.136	− 4.527	0.000
Emotional attitudes	− 0.127	− 4.211	0.000
Actual interactions	0.260	8.852	0.000
Animacy			
Social/future implications	− 0.095	− 3.129	0.002
Emotional attitudes	− 0.111	− 3.667	0.000
Actual interactions	0.199	6.695	0.000

scales should show no more reliability than a seven-point scale [16].

Regarding the comparison of the robots, we found a good sensitivity to dichotomize the stimuli with different patterns on each dimension arguing for complementarity between dimensions. As hypothesized, the Animacy dimension followed the ABOT database norm. Interestingly, the attribution of high cognitive capacities in the Agency dimension seems correlated to the human-like shape level of robots. Finally, Social and Disturbance traits presented opposed pattern. Therefore, it seems that all dimensions do not rely on the same vector of attribution but converge in a general dimension that is the anthropomorphic attribution.

Finally, the NARS make it possible to validate the external reliability of the questionnaire dimensions as positive attribution was perfectly predicted by positive attitudes (actual interactions dimension) while negative attribution was perfectly predicted by negative attitudes towards robots (social/future implications, emotional attitudes dimensions).

3.5 Study 4: Scale Validation in Real-World HRI

The first experiment makes it possible to define a suitable matrix scale to evaluate attribution to various range of robots on the four factors that are Sociability, Agency, Disturbance, and Animacy. The second experiment was designed to evaluate how the new questionnaire could account for the change of perception of a robot after a social interaction observation and assesses for the psychometric validity of the scale. The third experiment used an external database to create a controlled test sampling and assesses for the sensitivity of the scale. The fourth (and final) study tested the reliability of the questionnaire in order to finally evaluate the perception of robots online in real-world HRI. Recent studies showed that interaction with a robot could influence the attribution of anthropomorphic traits [112, 113]. We reinterpreted Spatola and colleagues' data (anthropomorphic questionnaires) in light of the present factorial structure [112]. Indeed, a social robot (i.e., a robot with social verbal interaction capacities) was seen with more uniquely human traits (e.g. warmth) and less mechanical traits (e.g. inertness) than a passive robot (i.e., a robot displaying the same physical movements than the social robot but without any social and verbal interaction). In line with the original study [112], regarding the present scale, the social robot interaction should elicit more sociability and agency because of the social nature of the interaction and the related inference associated with the robot that energizes a mentalization process in the observer [13, 33]. Also, more Animacy attribution should be made because, compared to a non-interactive robot, the social robot should be seen as more autonomous and with more technical and technological capacities [42]. Finally, the robot in the social interaction condition should elicit less Disturbance compared to the simple observation, as it should enhance a bonding feeling [72]. Finally, we expected to find the same four factors psychometric construct than in study 1 in a real HRI context.

3.5.1 Method

Participants Participants were 81 students from Clermont-Auvergne university ($M_{\text{age}} = 19.33$, SD 2.42, 65% males, 35% females) recruited in exchange of credit class.

Material and Procedure In a «non-social robot» condition (n = 40), participants were asked to give their opinion on the appearance of a physically present but passive robot. In the «social robot» condition (n = 41), participants were asked to interact verbally with the same robot that was controlled at distance by a human operator (without their knowledge) in a "Wizard of Oz paradigm" paradigm [49]. In both conditions, the robot had exactly the same preprogrammed movements. The robot was a 1-m MeccanoidG15KS humanoid that has already been used in similar experiments [112–114]. The operator was using two smartphones for the control of the robot's gestures and speech (by selecting pre-established conversational scripts) in a coherent way. This verbal interaction was set to encourage anthropomorphic inferences and familiarity towards the robot [101]. The interaction always followed the same pre-established script (see supplementary material), the operator only had to choose when to launch a given sequence. After the interaction, a French version of the scale was presented to the participants.[6] They had to judge to what extent the traits in the scale corresponded to the robot being present on a 7-point Likert scale 1 "not at all" to 7 "totally" in a paper-pen format. Participants made their judgments on a computer. Items were randomized to ensure the reliability of factors as not dependent on a semantic congruency effect order. Finally, in both conditions the cover story was to use their judgment to provide data for projects with roboticists, none of the participants declared any doubt about the purpose of the experiment during the debriefing. We also asked whether participants could have been disturbed by the interaction, all responses were negative.

3.5.2 Results

Structural Validity In order to conduct a confirmatory factor analysis we checked the Bartlett's sphericity test to ensure inter-item correlation ($\chi^2 = 812{,}52$, $ddl = 120$, $p < .001$) and the Kaiser–Meyer–Olkin Indice ($KMO = .84$) for the sample adequacy [9, 30, 60]. We used the same structural model (Fig. 3) using a variance–covariance matrix with maximum likelihood (ML) estimation. Table 15 shows the recommended model-fit [103] indices as well as the recommended thresholds.

Table 16 presents the non-standardized estimates for each item. All items were significantly associated with their respective factor (all $p_s < .001$).

Experimental Manipulation We conducted a multivariate ANOVA including all factors as DVs and the type of interaction with the robot as independent variable (non-social robot vs. social robot). Results showed that participants attributed significantly higher Sociability [$F(1{,}80) = 10.83$, $p = .001$,

[6] To translate the questionnaire from English to French we processed as follow. First, in a forward translation, two bilingual translators have translated the questionnaire into French. As recommended one translator was aware of the purpose of the questionnaire while the second one was naïve [110, 116]. The initial translation was independently back-translated in a backward process and we conducted a pre-test on the questionnaire to ensure psychometric reliability.

Table 15 Confirmatory model fit indices. χ^2/df the ratio of Chi square to degree of freedom; GFI the goodness-of-fit-index; AGFI the adjusted goodness-of-fit; NFI the normalized fit index, CFI the comparative fit index; RMSR the root mean square residual

	Recommended value	Values obtained
χ^2/df	$\leq$ 3.00	1.01
GFI	$\geq$ 0.90	0.87
AGFI	$\geq$ 0.80	0.80
NFI	$\geq$ 0.90	0.89
CFI	$\geq$ 0.90	0.99
TLI	$\geq$ 0.90	0.99
RMSEA	$\leq$ 0.08	0.01
SRMR	$\leq$ 0.08	0.08

$\eta_p^2 = .12$], Animacy [$F(1{,}80) = 5.70$, $p = .019$, $\eta_p^2 = .07$] and Agency [$F(1{,}80) = 6.21$, $p = .015$, $\eta_p^2 = .07$] traits to the robot in the social interaction condition compared to the non-social one. In addition, less Disturbance traits were associated to the robot in the social interaction condition [$F(1{,}80) = 13.58$, $p < .001$, $\eta_p^2 = .15$].

3.5.3 Discussion Study 3

First, this study aimed to replicate the psychometric construct of studies 1, 2, and 3 based real-world HRI data [112]. Results confirmed that the four-dimension pattern matrix is reliable according to Cronbach's alpha [19, 62].

Second, in agreement with our hypotheses, we found higher Sociability, Agency, and Animacy attribution in parallel to less Disturbance when evaluating the social robot compared to the non-social one. These results are in line with previous studies using the same methodology in which participants attributed more anthropomorphic characteristics to the robots after a social-compared to a non-social interaction [112–115]. However, comparing to these previous results, the present scale seems more sensitive than those used in the above-mentioned study. Interestingly, considering the positive but relative correlation of the four dimensions and the significance of each regarding the experimental manipulation, Sociability, Agency, Animacy, and Disturbance seem to reliably measure different components of anthropomorphism.

4 Presentation of the Human–Robot Interaction Evaluation Scale

To use this scale simply present the items on a 7-points Likert scale with the following instruction:

Using the scale provided, how closely are the words below associated with the [robot stimuli to evaluate]? From 1 "not at all" to 7 "totally".

Table 16 CFA non-standardized estimates

Items		Factor	Estimate	S.E.	t value	*p* value
Warm	←	Sociability	0.706	0.081	8.666	<.001
Trustworthy	←	Sociability	0.615	0.083	7.435	<.001
Likeable	←	Sociability	0.765	0.091	8.4	<.001
Friendly	←	Sociability	0.754	0.088	8.555	<.001
Scary	←	Disturbance	0.982	0.116	8.458	<.001
Creepy	←	Disturbance	1.064	0.111	9.565	<.001
Uncanny	←	Disturbance	0.963	0.111	8.649	<.001
Weird	←	Disturbance	1.048	0.124	8.468	<.001
Intelligent	←	Agency	0.804	0.098	8.239	<.001
Rational	←	Agency	0.502	0.102	4.905	<.001
Intentional	←	Agency	0.947	0.107	8.815	<.001
Conscious	←	Agency	0.618	0.106	5.833	<.001
Humanlike	←	Animacy	0.789	0.099	7.98	<.001
Alive	←	Animacy	0.917	0.105	8.689	<.001
Natural	←	Animacy	0.85	0.104	8.138	<.001
Real	←	Animacy	0.934	0.11	8.472	<.001

Items	Factor
Warm	Sociability
Likeable	Sociability
Trustworthy	Sociability
Friendly	Sociability
Alive	Animacy
Natural	Animacy
Real	Animacy
Human-like	Animacy
Self-reliant	Agency
Rational	Agency
Intentional	Agency
Intelligent	Agency
Creepy	Disturbance
Scary	Disturbance
Uncanny	Disturbance
Weird	Disturbance

We highly recommend randomization, at least of the factors, so not all participants evaluate each item in the same order which could, potentially, result in semantic bias. The structure of the scale holds with higher ranging scales however one should take into account the number of response possibilities when planning the experiment and the number of participants to not artificially increase interindividual variability. Therefore, we recommend the 7-points Likert scale. To analyze the score of participants we recommend considering the four dimensions separately while checking for collinearity. In the current state of knowledge, it is indeed difficult to consider each dimension as illustrating evaluation processes at the same levels. It is likely that one dimension may precede and therefore condition a subsequent evaluation process on another dimension.

5 General Limits

How humans consider and perceive robots is a complex topic as we still do not know and understand all factors implied. The present scale is, thus, dependent on actual conceptualization of HRI that is a simplified interaction compared to human–human interaction. Indeed, while there is basic inter-individual perception and evaluation, a broad range of socio-cognitive factors interact to define how we will consider and perceive others (e.g., conformism, intra-group bias) but also individual factors such as the feeling of loneliness, need for control, etc. All these determinants could affect how we perceive robots. Thus, to understand HRI, researchers must foster structuring perspectives more than a unitary approach.

Considering a central individual factor, in pretest and study 3, the sample was principally female and several studies demonstrated a gender effect on attitudes toward robots [31, 36, 89]. For instance, individuals experienced more psychological closeness to a same-sex robot than toward a robot of the opposite sex and most people report a preference for human avatars that matched their gender [92]. This gender effect could affect the anthropomorphic attribution and thus the result of the scale. However, this bias does not seem to impair the structure in perspective of the other studies presented in the manuscript. Still comparing the response tendency according to dispositional factors, as mentioned above, seem of great interest for social robotics.

Finally, scales aim to measure attitudes toward a stimulus or phenomenon. However, there are two forms of attitudes: explicit and implicit [34]. Explicit attitudes operate on a conscious level and are generally measured through self-report measures (e.g. questionnaires) while implicit attitudes often rely on the unconscious and automatic processes measured, e.g. through reaction time paradigms (e.g. implicit association test) [23]. In other words, implicit attitudes do not require a person's awareness or reflexive processing. These two forms of attitudes are sometimes related [59], however, implicit attitudes are showed as better predictors of future intention and behavior, especially in the inter-group relationship [85, 122]. A considerable amount of research suggests that attitudes toward others in an intergroup relationship are often based on implicit perceptions of these groups [3, 47]. In the context of human–robot interaction (HRI), robots may be seen as a group (i.e., as "non-human machines"). As in other social cognitive constructs, attitudes toward robots naturally arise from both conscious and unconscious processes [80, 118], so a combination of the proposed scale with implicit measures, e.g. heart-rate variability, reaction time, eye movements, skin conductance/-resistance, or EEG-based measures is desirable.

6 Conclusion

In order to improve our understanding of human–robot interactions, it is of prime importance to produce reliable tools to evaluate how we perceive these new artificial agents that we aim to integrate into our society, especially if we aim to use these agents as experimental tools to study human cognition [10, 14, 132]. In this article, we propose a new composite questionnaire to evaluate how people perceive robots and attribute human characteristics to them. The scale ranges from basic to uniquely human traits and measures the perception of others based on various state-of-the-art scales and psychological theories such as de-humanization. Considering the composite structures of robot evaluation is not trivial as, in interpersonal human behaviors, consequences of the attribution of human traits (or the opposite) is predictive of attitudes [33, 124] but also symptomatic of the form of the social evaluation process [51, 69]. With regard to robots, to attribute them to intentional traits, for example, is relied on the recognition of a form of individuality that relies on the same neural pathway as human–human interaction [13, 99]. Therefore the question is not to investigate if we anthropomorphize robot per se as it seems a default state but to what extent and what are the conditions for such a process to increase or decrease [124].

Regarding the evolution of social robotics, it is relevant to continuously improve and develop theories and tools to better envisage and evaluate the more and more complex nature of future human–robot relationships, especially with regard to social and psychological attributions. Based on intergroup psychological constructs and processes, the Human–Robot Interaction Evaluation Scale (HRIES) contributes to this interdisciplinary work by extending the evaluation dimension of robots essentially in real social HRI situations. The reliability and validity of the proposed scale are evaluated and confirmed in four different types of user studies, including different complexity-levels in their experimental design as used in the HRI community, ranging from online surveys over video-based studies, up to real-world HRI experiments.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical Approval This study was carried out in accordance with the provisions of the World Medical Association Declaration of Helsinki. Data were handle in accordance with the provisions of the General Data Protection Regulation (EU) 2016/679.

References

1. Anderson ML (2005) Why is AI so scary? Artif Intell 169(2):201–208. https://doi.org/10.1016/j.artint.2005.10.008
2. Bacher J, Wenzig K, Vogler M (2004) SPSS twostep cluster—a first evaluation. Univ Erlangen-Nürnberg 1(1):1–20
3. Banaji MR, Hardin C, Rothman AJ (1993) Implicit stereotyping in person judgment. J Pers Soc Psychol 65(2):272–281. https://doi.org/10.1037/0022-3514.65.2.272
4. Bartneck C, Kulić D, Croft E, Zoghbi S (2009) Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Int J Soc Robot 1:71–81. https://doi.org/10.1007/s12369-008-0001-3
5. Bollen KA (1989) Structural equations with latent variables. https://doi.org/10.1002/9781118619179
6. Bowling A, Windsor J (2008) The effects of question order and response-choice on self-rated health status in the English Longitudinal Study of Ageing (ELSA). J Epidemiol Community Health 62(1):81–85. https://doi.org/10.1136/jech.2006.058214
7. Burleigh TJ, Schoenherr JR, Lacroix GL (2013) Does the uncanny valley exist? An empirical test of the relationship between eeriness and the human likeness of digitally created faces. Comput Hum Behav 29(3):759–771. https://doi.org/10.1016/j.chb.2012.11.021
8. Carpinella CM, Wyman AB, Perez MA, Stroessner SJ (2017) The Robotic Social Attributes Scale (RoSAS): development and validation. ACM/IEEE Int Conf Hum Robot Interact Part F1271:254–262. https://doi.org/10.1145/2909824.3020208
9. Cerny BA, Kaiser HF (1977) A study of a measure of sampling adequacy for factor-analytic correlation matrices. Multivar Behav Res 12(1):43–47. https://doi.org/10.1207/s15327906mbr1201_3
10. Chaminade T, Cheng G (2009) Social cognitive neuroscience and humanoid robotics. J Physiol Paris 103(3–5):286–295. https://doi.org/10.1016/j.jphysparis.2009.08.011
11. Chaminade T, Franklin DW, Oztop E, Cheng G (2005) Motor interference between humans and humanoid robots: effect of biological and artificial motion. In: Proceedings of 2005 4th IEEE international conference on development and learning, 2005, pp 96–101. https://doi.org/10.1109/DEVLRN.2005.1490951
12. Chaminade T, Hodgins J, Kawato M (2007) Anthropomorphism influences perception of computer–animated characters' actions. Soc Cogn Affect Neurosci 2(3):206–216. https://doi.org/10.1093/scan/nsm017
13. Chaminade T, Rosset D, Da Fonseca D, Nazarian B, Lutcher E, Cheng G, Deruelle C (2012) How do we think machines think? An fMRI study of alleged competition with an artificial intelligence. Front Hum Neurosci. https://doi.org/10.3389/fnhum.2012.00103
14. Cheng G (2014) Humanoid robotics and neuroscience: science, engineering, and society. https://doi.org/10.1201/b17949
15. Choi BCK, Pak AWP (2005) Peer reviewed: a catalog of biases in questionnaires. Prevent Chronic Dis 2:A13
16. Cicchetti DV, Shoinralter D, Tyrer PJ (1985) The effect of number of rating scale categories on levels of interrater reliability: a Monte Carlo investigation. Appl Psychol Meas. https://doi.org/10.1177/014662168500900103
17. Cohen RJ, Swerdlik ME (2013) Psychological testing and assessment: an introduction to tests and measurement, 9th edn. McGrawHill, New York
18. Colman AM, Norris CE, Preston CC (1997) Comparing rating scales of different lengths: equivalence of scores from 5-point and 7-point scales. Psychol Rep. https://doi.org/10.2466/pr0.1997.80.2.355
19. Cronbach LJ (1951) Coefficient alpha and the internal structure of tests. Psychometrika 16(3):297–334. https://doi.org/10.1007/BF02310555
20. Darling K (2012) Extending legal rights to social robots. SSRN Electron J. https://doi.org/10.2139/ssrn.2044797
21. Darling K (2017) "Who's johnny?" Anthropomorphic framing in human–robot: interaction, integration, and policy. In: Robot ethics 2.0: from autonomous cars to artificial intelligence. https://doi.org/10.1093/oso/9780190652951.003.0012
22. De Graaf MMA, Ben Allouch S (2013) Exploring influencing variables for the acceptance of social robots. Robot Auton Syst 61(12):1476–1486. https://doi.org/10.1016/j.robot.2013.07.007
23. De Houwer J, Teige-Mocigemba S, Spruyt A, Moors A (2009) Implicit measures: a normative analysis and review. Psychol Bull 135(3):347–368. https://doi.org/10.1037/a0014211
24. de Leon SP, Lara-Muñoz C, Feinstein AR, Wells CK (2004) A comparison of three rating scales for measuring subjective phenomena in clinical research. Arch Med Res. https://doi.org/10.1016/j.arcmed.2003.07.009
25. De Winter JCF, Dodou D (2016) Common factor analysis versus principal component analysis: a comparison of loadings by means of simulations. Commun Stat Simul Comput 45(1):299–321. https://doi.org/10.1080/03610918.2013.862274
26. Dennett D (2009) Intentional systems theory. Oxf Handb Philos Mind 68(4):87–106. https://doi.org/10.1093/oxfordhb/9780199262618.003.0020

27. Dennett DC (1988) Précis of the intentional stance. Behav Brain Sci 11(3):495–505. https://doi.org/10.1017/S0140525X00058611
28. Diab LN (1965) Some limitations of existing scales in the measurement of social attitudes. Psychol Rep 17(2):427–430. https://doi.org/10.2466/pr0.1965.17.2.427
29. Dupree CH, Fiske ST (2017) Universal dimensions of social signals: warmth and competence. In: Social signal processing, pp 23–33. https://doi.org/10.1017/9781316676202.003
30. Dziuban CD, Shirkey EC (1974) When is a correlation matrix appropriate for factor analysis? Some decision rules. Psychol Bull 81(6):358–361. https://doi.org/10.1037/h0036316
31. Echterhoff G, Bohner G, Siebler F (2006) Social robotics and human–machine interaction: current research and relevance for social psychology. Zeitschrift Fuer Sozialpsychologie. https://doi.org/10.1024/0044-3514.37.4.219
32. Ehrlich SK, Cheng G (2018) Human-agent co-adaptation using error-related potentials. J Neural Eng. https://doi.org/10.1088/1741-2552/aae069
33. Epley N, Waytz A, Cacioppo JT (2007) On seeing human: a three-factor theory of anthropomorphism. Psychol Rev 114(4):864–886. https://doi.org/10.1037/0033-295X.114.4.864
34. Evans JSBT (2008) Dual-processing accounts of reasoning, judgment, and social cognition. Annu Rev Psychol 59(1):255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629
35. Eyssel F, Kuchenbrandt D (2012) Social categorization of social robots: anthropomorphism as a function of robot group membership. Br J Soc Psychol 51(4):724–731. https://doi.org/10.1111/j.2044-8309.2011.02082.x
36. Eyssel F, Kuchenbrandt D, Hegel F, De Ruiter L (2012) Activating elicited agent knowledge: how robot and user features shape the perception of social robots. In: Proceedings—IEEE international workshop on robot and human interactive communication. https://doi.org/10.1109/ROMAN.2012.6343858
37. Fayers P (2004) Item response theory for psychologists. Qual Life Res. https://doi.org/10.1023/b:qure.0000021503.45367.f2
38. Fink J (2012) Anthropomorphism and human likeness in the design of robots and human–robot interaction. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics). https://doi.org/10.1007/978-3-642-34103-8_20
39. Finn RH (1972) Effects of some variations in rating scale characteristics on the means and reliabilities of ratings. Educ Psychol Meas. https://doi.org/10.1177/001316447203200203
40. Fiske ST, Cuddy AJC, Glick P (2007) Universal dimensions of social cognition: warmth and competence. Trends Cogn Sci 11(2):77–83. https://doi.org/10.1016/j.tics.2006.11.005
41. Fiske ST, Neuberg SL (1990) A continuum of impression formation, from category-based to individuating processes: influences of information and motivation on attention and interpretation. Adv Exp Soc Psychol 23(4):1–74. https://doi.org/10.1016/S0065-2601(08)60317-2
42. Foster ME, Gaschler A, Giuliani M, Isard A, Pateraki M, Petrick RPA (2012) Two people walk into a bar : dynamic multi-party social interaction with a robot agent. ICMI. https://doi.org/10.1145/2388676.2388680
43. Friedman JH (1991) Multivariate adaptive regression splines. Ann Stat. https://doi.org/10.1214/aos/1176347963
44. Funke F, Reips UD (2012) Why semantic differentials in web-based research should be made from visual analogue scales and not from 5-point scales. Field Methods. https://doi.org/10.1177/1525822X12444061
45. Galesic M, Bosnjak M (2009) Effects of questionnaire length on participation and indicators of response quality in a web survey. Public Opin Q. https://doi.org/10.1093/poq/nfp031
46. Gorsuch RL (1990) Common factor analysis versus component analysis: some well and little known facts. Multivar Behav Res 25(1):33–39. https://doi.org/10.1207/s15327906mbr2501_3
47. Greenwald AG, Banaji MR (1995) Implicit social cognition: attitudes, self-esteem, and stereotypes. Psychol Rev 102(1):4–27. https://doi.org/10.1037/0033-295X.102.1.4
48. Hair JF, Black WC, Babin BJ, Anderson RE (2010) Multivariate data analysis. In: Exploratory data analysis in business and economics, 7th edn. Prentice Hall. https://doi.org/10.1007/978-3-319-01517-0_3
49. Hanington B, Martin B (2012) Universal methods of design: 100 ways to research complex problems, develop innovative ideas, and design effective solutions. In: Choice reviews online, vol 49. https://doi.org/10.5860/CHOICE.49-5403
50. Haslam N (2006) Dehumanization: an integrative review. Personal Soc Psychol Rev 10(3):252–264. https://doi.org/10.1207/s15327957pspr1003_4
51. Haslam N, Loughnan S (2012) Prejudice and dehumanization. Beyond Prejudice. https://doi.org/10.1017/cbo9781139022736.006
52. Haslam N, Loughnan S (2014) Dehumanization and infrahumanization. Annu Rev Psychol 65(1):399–423. https://doi.org/10.1146/annurev-psych-010213-115045
53. Heider F, Simmel M (1950) An experimental study of apparent behavior. Jpn J Psychol 20(2):67–74. https://doi.org/10.4992/jjpsy.20.2_67
54. Heil J, Heil J (2019) The intentional stance. Philos Mind. https://doi.org/10.4324/9780429506994-9
55. Heise DR (1969) Some methodological issues in semantic differential research. Psychol Bull. https://doi.org/10.1037/h0028448
56. Heise D (1970) The semantic differential and attitude research. In: Summers GF (ed) Attitude measurement. Rand McNally & Co, Chicago, pp 235–253
57. Hendrickson AE, White PO (1964) Promax: a quick method for rotation to oblique simple structure. Br J Stat Psychol. https://doi.org/10.1111/j.2044-8317.1964.tb00244.x
58. Ho CC, MacDorman KF (2010) Revisiting the uncanny valley theory: developing and validating an alternative to the Godspeed indices. Comput Hum Behav 26(6):1508–1518. https://doi.org/10.1016/j.chb.2010.05.015
59. Hofmann W, Gawronski B, Gschwendner T, Le H, Schmitt M (2005) A meta-analysis on the correlation between the implicit association test and explicit self-report measures. Pers Soc Psychol Bull 31:1369–1385. https://doi.org/10.1177/0146167205275613
60. IBM (2011) IBM Kaiser–Meyer–Olkin measure for identity correlation matrix. J R Stat Soc. http://www-01.ibm.com/support/docview.wss?uid=swg21479963
61. Jackson DL, Gillaspy JA, Purc-Stephenson R (2009) Reporting practices in confirmatory factor analysis: an overview and some recommendations. Psychol Methods. https://doi.org/10.1037/a0014694
62. Brown JD (2002) The Cronbach alpha reliability estimate. Shiken JALT Test Eval SIG Newsl 6(1):17–18
63. Joyce CRB, Zutshi DW, Hrubes V, Mason RM (1975) Comparison of fixed interval and visual analogue scales for rating chronic pain. Eur J Clin Pharmacol. https://doi.org/10.1007/BF00562315
64. Kätsyri J, Förger K, Mäkäräinen M, Takala T (2015) A review of empirical evidence on different uncanny valley hypotheses: support for perceptual mismatch as one road to the valley of eeriness. Front Psychol. https://doi.org/10.3389/fpsyg.2015.00390
65. Kelley HH (1967) Attribution in social psychology. In: Jones EE, Kanouse DE, Kelley HH et al (eds) Nebraska Symposium on Motivation. Lawrence Erlbaum Associates Inc, Hillsdale, NJ, US, pp 192–238

66. Kiesler S, Powers A, Fussell SR, Torrey C (2008) Anthropomorphic interactions with a robot and robot–like agent. Soc Cogn 26(2):169–181. https://doi.org/10.1521/soco.2008.26.2.169
67. Krosnick JA, Boninger DS, Chuang YC, Berent MK, Carnot CG (1993) Attitude strength: one construct or many related constructs? J Pers Soc Psychol. https://doi.org/10.1037/0022-3514.65.6.1132
68. Kteily N, Bruneau E, Waytz A, Cotterill S (2015) The ascent of man: theoretical and empirical evidence for blatant dehumanization. J Pers Soc Psychol 109(5):901–931. https://doi.org/10.1037/pspp0000048
69. Kteily N, Hodson G, Bruneau E (2016) They see us as less than human: metadehumanization predicts intergroup conflict via reciprocal dehumanization. J Pers Soc Psychol. https://doi.org/10.1037/pspa0000044
70. Kuchenbrandt D, Eyssel F, Bobinger S, Neufeld M (2013) When a robot's group membership matters: anthropomorphization of robots as a function of social categorization. Int J Soc Robot 5(3):409–417. https://doi.org/10.1007/s12369-013-0197-8
71. Kühnlenz B, Maximilian E, Marcel K, Zhi-Qiao W, Julian W, Kolja K (2018) Impact of trajectory profiles on user stress in close human–robot interaction. At Automatisierungstechnik 66:483. https://doi.org/10.1515/auto-2018-0004
72. Kühnlenz B, Sosnowski S, Buß M, Wollherr D, Kühnlenz K, Buss M (2013) Increasing helpfulness towards a robot by emotional adaption to the user. Int J Soc Robot. https://doi.org/10.1007/s12369-013-0182-2
73. Lee S, Schwarz N (2014) Question context and priming meaning of health: effect on differences in self-rated health between Hispanics and non-Hispanic Whites. Am J Public Health. https://doi.org/10.2105/AJPH.2012.301055
74. Lewis JR (1993) Multipoint scales: mean and median differences and observed significance levels. Int J Hum Comput Interact. https://doi.org/10.1080/10447319309526075
75. Lindwall M, Barkoukis V, Grano C, Lucidi F, Raudsepp L, Liukkonen J, Thgersen-Ntoumani C (2012) Method effects: the problem with negatively versus positively keyed items. J Pers Assess. https://doi.org/10.1080/00223891.2011.645936
76. Lopes PN, Salovey P, Côté S, Beers M (2005) Emotion regulation abilities and the quality of social interaction. Emotion 5:113–118. https://doi.org/10.1037/1528-3542.5.1.113
77. Lozano LM, García-Cueto E, Muñiz J (2008) Effect of the number of response categories on the reliability and validity of rating scales. Methodology. https://doi.org/10.1027/1614-2241.4.2.73
78. MacDorman K (2006) Subjective ratings of robot video clips for human likeness, familiarity, and eeriness: an exploration of the uncanny valley. In: ICCS/CogSci-2006 long symposium: toward …. https://doi.org/10.1093/scan/nsr025
79. MacDorman KF, Chattopadhyay D (2016) Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. Cognition. https://doi.org/10.1016/j.cognition.2015.09.019
80. MacDorman KF, Vasudevan SK, Ho CC (2009) Does Japan really have robot mania? Comparing attitudes by implicit and explicit measures. AI & Soc 23(4):485–510. https://doi.org/10.1007/s00146-008-0181-2
81. Marchesi S, Ghiglino D, Ciardo F, Perez-Osorio J, Baykara E, Wykowska A (2019) Do we adopt the intentional stance toward humanoid robots? Front Psychol. https://doi.org/10.3389/fpsyg.2019.00450
82. Mathur MB, Reichling DB (2016) Navigating a social world with robot partners: a quantitative cartography of the Uncanny Valley. Cognition 146:22–32. https://doi.org/10.1016/j.cognition.2015.09.008
83. Mavletova A (2013) Data quality in PC and mobile web surveys. Soc Sci Comput Rev. https://doi.org/10.1177/0894439313485201
84. Maxwell AE, Harman HH (2006) Modern factor analysis. J R Stat Soc Ser A (Gen). https://doi.org/10.2307/2343736
85. McConnell AR, Leibold JM (2001) Relations among the implicit association test, discriminatory behavior, and explicit measures of racial attitudes. J Exp Soc Psychol 37(5):435–442. https://doi.org/10.1006/jesp.2000.1470
86. Meade AW, Craig SB (2012) Identifying careless responses in survey data. Psychol Methods. https://doi.org/10.1037/a0028085
87. Mishra M (2016) Confirmatory factor analysis (CFA) as an analytical technique to assess measurement error in survey research. Paradigm. https://doi.org/10.1177/0971890716672933
88. Mori M, MacDorman KF, Kageki N (2012) The uncanny valley. IEEE Robot Autom Mag 19(2):98–100. https://doi.org/10.1109/MRA.2012.2192811
89. Nomura T, Kanda T, Suzuki T (2006) Experimental investigation into influence of negative attitudes toward robots on human–robot interaction. AI Soc. https://doi.org/10.1007/s00146-005-0012-7
90. Nomura T, Suzuki T, Kanda T, Kato K (2006) Measurement of anxiety toward robots. In: Proceedings—IEEE international workshop on robot and human interactive communication, pp 372–377. https://doi.org/10.1109/ROMAN.2006.314462
91. Nomura T, Suzuki T, Kanda T, Kato K (2006) Measurement of negative attitudes toward robots. Interact Stud. https://doi.org/10.1075/is.7.3.14nom
92. Nowak KL, Rauh C (2005) The influence of the avatar on online perceptions of anthropomorphism, androgyny, credibility, homophily, and attraction. J Comput Mediat Commun. https://doi.org/10.1111/j.1083-6101.2006.tb00308.x
93. Oxford English Dictionary (2017) Oxford English dictionary online. Oxford University Press, Oxford, UK
94. Oztop E, Chaminade T, Franklin DW (2004) Human-humanoid interaction: is a humanoid robot perceived as a human? IEEE/RAS Int Conf Humanoid Robots 2(4):830–841. https://doi.org/10.1109/ICHR.2004.1442688
95. Pérez-Osorio J, Wykowska A (2019) Adopting the intentional stance toward natural and artificial agents. Philos Psychol. https://doi.org/10.31234/osf.io/t7dwg
96. Phillips E, Zhao X, Ullman D, Malle BF (2018) What is human-like? Decomposing robots' human-like appearance using the anthropomorphic roBOT (ABOT) database. ACM/IEEE Int Conf Hum Robot Interact. https://doi.org/10.1145/3171221.3171268
97. Preston CC, Colman AM (2000) Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. Acta Physiol (Oxf). https://doi.org/10.1016/S0001-6918(99)00050-5
98. Ramsay JO (1973) The effect of number of categories in rating scales on precision of estimation of scale values. Psychometrika. https://doi.org/10.1007/BF02291492
99. Rauchbauer B, Nazarian B, Bourhis M, Ochs M, Prévot L, Chaminade T (2019) Brain activity during reciprocal social interaction investigated using conversational robots as control condition. Philos Trans R Soc B Biol Sci. https://doi.org/10.1098/rstb.2018.0033
100. Roszkowski MJ, Soven M (2010) Shifting gears: consequences of including two negatively worded items in the middle of a positively worded questionnaire. Assess Eval High Educ. https://doi.org/10.1080/02602930802618344
101. Salem M, Eyssel F, Rohlfing K, Kopp S, Joublin F (2013) To err is human(-like): effects of robot gesture on perceived anthropomorphism and likability. Int J Soc Robot 5(3):313–323. https://doi.org/10.1007/s12369-013-0196-9
102. Saygin AP, Chaminade T, Ishiguro H, Driver J, Frith C (2012) The thing that should not be: predictive coding and the uncanny

valley in perceiving human and humanoid robot actions. Soc Cogn Affect Neurosci 7(4):413–422. https://doi.org/10.1093/scan/nsr025

103. Schermelleh-Engel K, Moosbrugger H, Müller H (2003) Evaluating the fit of structural equation models: tests of significance and descriptive goodness-of-fit measures. Methods Psychol Res Online 8:23–74

104. Schriesheim CA, Hill KD (1981) Controlling acquiescence response bias by item reversals: the effect on questionnaire validity. Educ Psychol Meas. https://doi.org/10.1177/001316448104100420

105. Schwarz N (1999) Self-reports: How the questions shape the answers. Am Psychol 54:93–105. https://doi.org/10.1037/0003-066x.54.2.93

106. Shepherdson P, Miller J (2014) Redundancy gain in semantic categorisation. Acta Physiol (Oxf). https://doi.org/10.1016/j.actpsy.2014.01.011

107. Shi J, Kashima Y, Loughnan S, Suitner C, Haslam N (2008) Subhuman, inhuman, and superhuman: contrasting humans with nonhumans in three cultures. Soc Cogn 26(2):248–258. https://doi.org/10.1521/soco.2008.26.2.248

108. Shoemaker S, Dennett D (1990) The intentional stance. J Philos 87(4):212. https://doi.org/10.2307/2026682

109. Snook SC, Gorsuch RL (1989) Component analysis versus common factor analysis: a Monte Carlo study. Psychol Bull 106(1):148–154. https://doi.org/10.1037/0033-2909.106.1.148

110. Sousa VD, Rojjanasrirat W (2011) Translation, adaptation and validation of instruments or scales for use in cross-cultural health care research: a clear and user-friendly guideline. J Eval Clin Pract. https://doi.org/10.1111/j.1365-2753.2010.01434.x

111. Spatola N (2019) L'homme et le robot, de l'anthropomorphisme à l'humanisation. Top Cogn Psychol 119:515–563

112. Spatola N, Belletier C, Chausse P, Augustinova M, Normand A, Barra V et al (2019) Improved cognitive control in presence of anthropomorphized robots. Int J Soc Robot 11(3):463–476. https://doi.org/10.1007/s12369-018-00511-w

113. Spatola N, Belletier C, Normand A, Chausse P, Monceau S, Augustinova M et al (2018) Not as bad as it seems: when the presence of a threatening humanoid robot improves human performance. Sci Robot 3(21):eaat5843. https://doi.org/10.1126/scirobotics.aat5843

114. Spatola N, Monceau S, Ferrand L (2019) Cognitive impact of social robots: how anthropomorphism boosts performance. IEEE Robot Autom Mag. https://doi.org/10.1109/MRA.2019.2928823

115. Spatola N, Santiago J, Beffara B, Mermillod M, Ferrand L, Ouellet M (2018) When the sad past is left: the mental metaphors between time, valence, and space. Front Psychol. https://doi.org/10.3389/fpsyg.2018.01019

116. Sperber AD (2004) Translation and validation of study instruments for cross-cultural research. Gastroenterology. https://doi.org/10.1053/j.gastro.2003.10.016

117. Stansbury JP, Ried LD, Velozo CA (2006) Unidimensionality and bandwidth in the center for epidemiologic studies depression (CES-D) scale. J Pers Assess. https://doi.org/10.1207/s15327752jpa8601_03

118. Sumioka H, Złotowski J, Nishio S, Eyssel F, Ishiguro H, Bartneck C (2018) Model of dual anthropomorphism: the relationship between the media equation effect and implicit anthropomorphism. Int J Soc Robot 10(5):701–714. https://doi.org/10.1007/s12369-018-0476-5

119. Sundar SS, Waddell TF, Jung EH (2016) The Hollywood robot syndrome: media effects on older adults' attitudes toward robots and adoption intentions. ACM/IEEE Int Conf Hum Robot Interact. https://doi.org/10.1109/HRI.2016.7451771

120. Symonds PM (1924) On the loss of reliability in ratings due to coarseness of the scale. J Exp Psychol. https://doi.org/10.1037/h0074469

121. Syrdal DS, Dautenhahn K, Koay KL, Walters ML (2009) The negative attitudes towards robots scale and reactions to robot behaviour in a live human–robot interaction study. In: Adaptive and emergent behaviour and complex systems—proceedings of the 23rd convention of the society for the study of artificial intelligence and simulation of behaviour, AISB 2009

122. Tetlock PE, Oswald FL, Mitchell G, Blanton H, Jaccard J (2013) Predicting ethnic and racial discrimination: a meta-analysis of IAT criterion studies. J Pers Soc Psychol 105(2):171–192. https://doi.org/10.1037/a0032734

123. Unwin A (2013) Discovering statistics using R by Andy Field, Jeremy Miles, Zoë Field. In: International statistical review, vol 81. https://doi.org/10.1111/insr.12011_21

124. Urquiza-Haas EG, Kotrschal K (2015) The mind behind anthropomorphic thinking: attribution of mental states to other species. Anim Behav 109:167–176. https://doi.org/10.1016/j.anbehav.2015.08.011

125. Vandenberg RJ (2006) Introduction: statistical and methodological myths and urban legends. Organ Res Methods. https://doi.org/10.1177/1094428105285506

126. Velicer WF, Jackson DN (1990) Component analysis versus common factor analysis: some issues in selecting an appropriate procedure. Multivar Behav Res 25(1):1–28. https://doi.org/10.1207/s15327906mbr2501_1

127. Waytz A, Morewedge CK, Epley N, Monteleone G, Gao JH, Cacioppo JT (2010) Making sense by making sentient: effectance motivation increases anthropomorphism. J Pers Soc Psychol 99(3):410–435. https://doi.org/10.1037/a0020240

128. Widaman KF (1993) Common factor analysis versus principal component analysis: differential bias in representing model parameters? Multivar Behav Res 28(3):263–311. https://doi.org/10.1207/s15327906mbr2803_1

129. Williams B, Onsman A, Brown T (2018) Exploratory factor analysis: a five-step guide for novices. Australas J Paramed 4:5. https://doi.org/10.33151/ajp.8.3.93

130. Wood P (2008) Confirmatory factor analysis for applied research. In: The American statistician, vol 62. https://doi.org/10.1198/tas.2008.s98

131. Worthington RL, Whittaker TA (2006) Scale development research: a content analysis and recommendations for best practices. Couns Psychol. https://doi.org/10.1177/0011000006288127

132. Wykowska A, Chaminade T, Cheng G (2016) Embodied artificial agents for understanding human social cognition. Philos Trans R Soc B Biol Sci. https://doi.org/10.1098/rstb.2015.0375

133. Xie Y, DeVellis RF (2006) Scale development: theory and applications. Contemp Sociol. https://doi.org/10.2307/2075704

134. Yamagishi T (2001) Trust as a form of social intelligence. In: Cook KS (ed) Trust in society. Russell Sage foundation series on trust, vol 2. Russell Sage Foundation, New York, NY, pp 121–147

135. Yang GZ, Bellingham J, Dupont PE, Fischer P, Floridi L, Full R et al (2018) The grand challenges of science robotics. Sci Robot 3(14):eaar7650. https://doi.org/10.1126/scirobotics.aar7650

136. Yarkin KL, Harvey JH, Bloxom BM (1981) Cognitive sets, attribution, and social interaction. J Pers Soc Psychol. https://doi.org/10.1037/0022-3514.41.2.243

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Nicolas Spatola is postdoctoral researcher at the Social Cognition in Human-Robot Interaction lab, Istituto Italiano di Tecnologia, Genoa, Italy. His research examines how the development of social robotics and Artificial Intelligence may impact individuals' cognition and society. Also, his studies include how Human–Robot Interaction may be perceived by humans in regard to cognitive and socio-cognitive processes. His research interests promote these questions in an interdisciplinary perspective from psychological sciences to economical sciences and social robotics.

Barbara Kühnlenz is professor of business psychology with focus on technical innovation at the Academic Center for Sciences and Humanities, Coburg University of Applied Sciences and Arts, Germany. Barbara holds a master's degree (Magister Artium) in Psycholinguistics and Social Psychology from Ludwig-Maximilians-University (LMU), and received her PhD degree in 2013 (summa cum laude) by the Electrical Engineering and Information Technology department at the Technical University of Munich (TUM). She performed her PhD at the Institute of Automatic Control Engineering (LSR) as a member of the interdisciplinary cluster of excellence 'Cognition for Technical Systems' (CoTeSys) and Institute for Advanced Studies (IAS) with a research focus on social Human–Robot Interaction (HRI). Barbara was postdoctoral researcher at the Institute for Cognitive Systems (ICS) from 2018 to 2019. Before, she was scientific coordinator of TechnologieAllianzOberfranken (TAO), from 2014 to 2017, and postdoctoral researcher and lecturer at the Electrical Engineering and Information Technology department at the University for Applied Sciences in Coburg from 2016 to 2019.

Gordon Cheng has made pioneering contributions in Humanoid Robotics, Neuroengineering, Artificial Intelligence for the past 20 years. He is Founder and Director of Institute for Cognitive Systems, Faculty of Electrical and Computer Engineering at Technical University of Munich, Munich/Germany. He is the coordinator of the Center of Competence Neuro-Engineering (2013-). He is the director of the Elite Master of Science program in Neuroengineering (MSNE) of the Elite Network of Bavaria (2016-). His research interests include humanoid robotics, cognitive systems, artificial robot skin, brain-machine interfaces, bio-mimetic of human vision, computational neuroscience of vision, action understanding, human-robot interaction, active vision and mobile robot navigation. He is the co-inventor of approximately 20 patents and author of approximately 300 technical publications, proceedings, editorials and book chapters. He has been named IEEE Fellow 2017 for "contributions in humanoid robotic systems and neurorobotics".