AI & SOCIETY
https://doi.org/10.1007/s00146-020-01114-8

ORIGINAL ARTICLE

Attribution of autonomy and its role in robotic language acquisition

Frank Förster[1,2] · Kaspar Althoefer[1]

Received: 6 March 2020 / Accepted: 2 November 2020

Abstract
The false attribution of autonomy and related concepts to artificial agents that lack the attributed levels of the respective characteristic is problematic in many ways. In this article, we contrast this view with a positive viewpoint that emphasizes the potential role of such false attributions in the context of robotic language acquisition. By adding emotional displays and congruent body behaviors to a child-like humanoid robot's behavioral repertoire, we were able to bring naïve human tutors to engage in so called intent interpretations. In developmental psychology, intent interpretations can be hypothesized to play a central role in the acquisition of emotion, volition, and similar autonomy-related words. The aforementioned experiments originally targeted the acquisition of linguistic negation. However, participants produced other affect- and motivation-related words with high frequencies too and, as a consequence, these entered the robot's active vocabulary. We will analyze participants' non-negative emotional and volitional speech and contrast it with participants' speech in a non-affective baseline scenario. Implications of these findings for robotic language acquisition in particular and artificial intelligence and robotics more generally will also be discussed.

Keywords Language acquisition · Developmental robotics · Artificial intelligence · Human–robot interaction

1 Introduction

Humans appear to have the general tendency to interpret events and processes of unknown origin in agent-centric ways (Levinson 1995). In other words, human intelligence is lopsided, with a preference for social explanations when confronted with effects and processes of unknown origin and causation. This 'cognitive skewness' may be a side effect of our intelligence being the intelligence of a highly social species. Beyond this strong tendency to assume social reasons and human or, at least, agentic causation, there is a second, but possibly related tendency: the attribution of psychological and biological qualities to objects and technical artifacts which they genuinely do not possess, or which they possess to a lesser degree than is attributed to them. This tendency seems to be especially strong when humans perceive a lack in meaningful connections to other humans as is the case when they feel lonely (Epley et al. 2008). In the context of this special issue, the most relevant qualities amongst these are attributed autonomy and agency. Scheutz (2011) has pointed out the danger of users falsely attributing emotions to social robots in that emotionally unidirectional may be detrimental for genuine reciprocal human relationships. As we will outline below, the attribution of emotion is one element of the attribution of agency more generally.

In this paper, we will conduct an analysis of robot-directed speech based on speech transcripts originating from a set of experiments that targeted the acquisition and grounding of linguistic negation (Förster et al. 2018, 2019). These recordings were analyzed in the past with an exclusive focus on negation words such as "no" or "don't", many of which can be linked to so called *intent interpretations* on the pragmatic level. These intent interpretations appear to be linguistic indicators of the attribution of autonomy or agency. While we focused exclusively on negative intent interpretations in the aforementioned studies, the current analysis extends the scope to include their positive counterparts.

✉ Frank Förster
frank.foerster@gmx.net

Kaspar Althoefer
k.althoefer@qmul.ac.uk

[1] Centre for Advanced Robotics @ Queen Mary (ARQ), Queen Mary University of London, Mile End Road, London E1 4NS, UK

[2] Adaptive Systems Research Group, School of Physics, Engineering and Computer Science, University of Hertfordshire, College Lane, Hatfield AL10 9AB, UK

Published online: 16 January 2021

One problem that we are facing, when trying to isolate linguistic evidence of attributions of autonomy or agency is a lack of consensus in the robotics community as to what precisely constitutes each of these two concepts. There are numerous features that are stated to be important for both concepts and, to make matters worse, having motivation or volition, being able to experience emotions, and the capacity to govern one's own actions are frequently mentioned as such central features.

1.1 Agency and autonomy: two overlapping concepts

Before discussing the ways in which perceptions of agency and autonomy are commonly quantified, it is opportune to first introduce the concepts themselves. While agency and autonomy are rarely discussed explicitly side-by-side, one rare occasion is provided by Scheutz (2011). Citing Buss (2002), he adopts a philosophical definition of being autonomous as "to be a law to oneself", including the statement that "autonomous agents are self-governing agents" (ibid.). Scheutz subsequently boils the relationship between autonomy and agency down to the formula "Autonomy + Mobility = Agency". Precisely contrary to the current authors' intuition, autonomy is here the more general concept and agency a sub concept thereof. Mobility, which we would have thought of as a crucial ingredient for autonomy, and which is frequently part of the relevant psychological scales, is made a feature of agency here that sets it apart from autonomy more generally. Brincker (2016), discussing the perception of agency and its potential role for interaction, seems to be more in agreement with the present authors' intuition in regarding agency as the more general of the two concepts—perceived autonomy is seen here as one of the "ingredients" of perceived agency. Of some relevance for our data analyses will be the fact that Brincker, referencing the enactivist notion of autopoiesis (Varela et al. 1974), links up perceived agency with the perception of an agent having some form of metabolism [see also Gahrn-Andersen (2020) for a discussion of perceived vs. naturalistic autonomy]. Brincker also relates perceived agency to the perceiving agents' third- and second-person view of the observed agent's affordance field.[1] Of particular interest for our purposes is her observation that spatially co-located agents, when entering a reciprocal interaction,[2] adopt a second-person view of the shared affordance field. This view is ought to provide them with a particularly rich access to aspects of each other's agency. In this context, emotional expressions play an important role within the "triangulating dynamics" that both agents engage in when negotiating each other's roles, interpretations, and when trying to align or differentiate their respective perspective on the world. Here then, we find a connection between the role of emotional displays and attribution of agency and autonomy.

1.2 Perceptions of agency and autonomy

In psychology, a popular tool for assessing the perception of some particular quality is the scale or questionnaire. When searching for scales that would assess the perception of agency and autonomy, respectively, we failed to find one definite scale that most researchers in human–robot interaction (HRI) would utilize. Instead, we encountered a number of scales with the help of which perceptions of either of the two concepts are ought to be quantified. Within all these scales, however, the respective factor, agency or autonomy, was only one amongst several others. In other words, we are unaware of any scale that would assess perceived autonomy or perceived agency exclusively and which would have been designed for this designated purpose.

Most scales for assessing perceived autonomy appear to support our naïve intuition that autonomy seems to be strongly linked to physical mobility. The *EmCorp* scale (Hoffmann et al. 2018) measures perceived autonomy with mobility-related items such as the item 'robot is able to autonomously navigate in space'. Similarly, Schaefer and collegues (2018) assess users' perception of the ARIBO driverless shuttle, a vehicle somewhat similar to a golf cart, using the question "the vehicle controls itself", where control in this case is naturally linked to navigation due to the nature of the autonomous vehicle. In addition, Papenmeier et al. (2019) have participants assess their mobile robot repeatedly in terms of its autonomy. Given that the main distinguishing features between the different presentations are head orientation and the direction of movement of the robot, it is to little surprise that also here perceived autonomy is tightly linked to features of mobility. Here, autonomy is assessed via the single semantic differential 'pilot-operated' vs. 'self-propelled'.

Further support for a close relationship between autonomy and mobility or movement is provided by Jochum et al. (2017). Rather than report experimental results, the authors

[1] Brincker's notion of affordances is largely based on that of Rietveld and Kiverstein (2014), who define them as "relations between aspects of a material environment and abilities available in a form of life". 'Form of life' is a term borrowed from Wittgenstein and is meant to emphasise that human action is always executed within specific sociocultural contexts.

[2] Not every interaction in human–robot interaction is necessarily reciprocal, especially if the robot lacks the capability to assess the human's intentions correctly, hence the additional qualification.

provide insights into the design of entertainment robots and emphasize the importance of keeping a robot constantly animated if the "illusion of autonomy" is to be maintained for the observer.

Support of a broader interpretation of perceived autonomy which is tied less tightly to the notion of mobility and movement is provided by the *person perception scale* (Rosenthal-von der Pütten et al. 2017). This scale is used in the former study to assess participants' impression of a humanoid robot with respect to likability, intelligence, and autonomy. Judging by the items that make up the factor *autonomy*, the presumed underlying concept is ostensibly more general than is the case in *EmCorp*, with mobility not playing a major role. The six items making up the factor *autonomy* in the person perception scale are *'(not) autonomous*', '*self-dependent*', '*responsible for its actions*', '*restricted in its abilities*', *'free'*, and '*self-determined*'.

In its more broader take on perceived autonomy, the person perception scale is remarkably similar to a construct reported in Epley et al. (2008) which was used to assess participants' *perceived agency* of technical gadgets such as a mobile alarm clock or a programmable pillow. To enable participants to rate these gadgets in terms of their perceived agency Epley and collegues provide them with the items '*has a mind of its own*', '*has intentions*', '*has free will*', '*has consciousness*', and '*experiences emotion*'. Finally, Kamide et al. (2013) provide another construct for perceived agency together with eight other factors such as perceptions of utility, familiarity, or repulsion. All nine factors make up the *PHIT-40* scale that was designed specifically for assessing perceptions of humanoid robots. Within PHIT-40, perceived agency is calculated based on the two Likert-scale items "The robot looks as if it has a heart" and "The robot looks as if it has its own will".

Summarily, and based on the ways in which agency and autonomy are assessed within the aforementioned scales, we cannot fail to observe that the notion of autonomy seems to vary halfways consistently with the nature of the to-be-assessed robot. If the robots in question bear a resemblance with non-autonomous vehicles such as cars, autonomy assessments tend to make reference to mobility and the locus of control (self- vs. other-controlled). If the robots in question bare more resemblance to the human shape, the constructs for assessing perceptions of autonomy are wider, and tend to make reference to high-level psychological constructs such as emotions, intentions, and volition. Moreover, this wider version of perceived autonomy bear a strong resemblance to equivalent constructs for perceived agency. While questionnaires, or self-reports, may be the most widely used tools to measure personal attitudes or subjective construals of technical artifacts in HRI, there are alternative approaches. One such approach is the analysis of participants' speech-in-interaction, which is used in the present study and which carries the advantage of "measuring" or indicating such attitudes at the time of the interaction, rather than relying on a posteriori self-assessments. However, prior to introducing the methodology, we need to introduce the concept of intent interpretations, as they may play a central role in linguistic expressions of attributions of emotion, volition, or motivation.

1.3 Intent interpretations and attributions of agency

The concept of intent interpretations as we use it here, originates from research in developmental psychology on language acquisition. Pea (1980), attributing the notion to (Ryan 1974), and focusing on their import in the context of early negation, describes with intent interpretations those utterances that adults use to linguistically express non-linguistic acts of rejection by the child. Such acts of rejection include headshakes or the throwing away of offered things amongst others. However, intent interpretations are meant to cover also utterances that "describe" or interpret positive expressions of intention produced by a linguistically incompetent child. In our previous work, we have adapted the concept of *intent interpretations* in a analysis of negative robot-directed speech and defined negative intent interpretations as "assertions in which the participant interprets (the robot's) intentional or motivational state utilizing lexical and/or grammatical negatives" (Förster 2018). Another important property of this type of utterance is that it identifies a child's (or robot's) motivational state in the here and now and thus contrasts with habitual preferences or motivational expressions in the past or future. Intent interpretations are, therefore, by definition, temporally tightly linked to co-occurring non-linguistic expressions of the target of interpretation.

When viewed as an interpersonal dyadic mechanism—a linguistically more competent speaker "spells out" a less competent speaker's intentions and feelings in direct response to these very bodily expressions—intent interpretations form an interesting contrast to the more established joint attentional frames (cf. Tomasello 2003). The latter are said to be one of the central mechanisms in human language acquisition where a more competent speaker, typically the parent, shares attention with the language learner, typically a child, with respect to some outside entity (Tomasello 2000). In other words, joint attentional frames are a triadic mechanism that necessarily involve a third xentity in addition to the two interactors, whereas intent interpretations are dyadic mechanisms which may or may not involve some external referent. In joint attentional frames the ultimate target of the teacher's utterance, the thing that the utterance is about, lies outside of the dyad, whereas intent interpretations are ultimately about the learner's emotion, motivation, or volition.

Based on the previous analyses of participants' negative speech documented in Förster et al. (2019), the high prevalence of negative intent interpretations within that that speech, and the link of these utterances to the robot's bodily display of motivation or affect, we believe that intent interpretations more generally—positive or negative forms—are a highly frequent audible expression of human attributions of agency or autonomy to an agent. While we also encountered other forms of negative speech that may provide similarly strong clues with respect to such attributions, negative intent interpretations were by far the most frequent type of this kind amongst negative utterances. It is for this reason that we expect many non-negative utterances deemed to indicate participants' attributions of motivation, and therefore agency, to fall in this category.

2 Methods

The results presented in this paper are based on a combination of lexical analysis and semantic–pragmatic categorization of human participants' speech transcripts. Analyzing speech gathered during an interaction between humans and robots to gain insights into both the participants' perception of the robot and the interaction more general is not entirely new (cf. Fischer et al. 2012). While in HRI, the use of post-experimental self-reports is more frequently used in assessing types of participant attributions; Fischer and colleagues show that, since participants' linguistic choices correlate with their conceptualization of the robot (cf Fischer 2011 and Fischer et al. 2011), they can be used to assess the latter.

In the present study, the relevant speech was originally gathered during two experiments that targeted the acquisition and grounding of linguistic negation (Förster et al. 2018, 2019). In addition to these two corpora, we will use another corpus of speech transcripts as a baseline which was gathered during an experiment that preceded the two negation experiments (Saunders et al. 2012). The latter focused on the robotic acquisition and lexical grounding of object labels and of words relating to object attributes such as color or size. While details on these experiments are provided in the aforementioned publications, we summarize these three studies in the next section to sketch the situational context within which this material was gathered. The move clarifies the overall motivation behind using affective displays for the purpose of language acquisition.

While the aforementioned publications from Förster et al. focused analytically on negation words, the analysis of the present study hones in on emotion and volition words as well as words and utterances that indicate attributions of autonomy.

2.1 Experiments on the acquisition of negation, object labels, and object attributes

The set of experiments constituting the basis for the speech corpora used within the following analysis combined methods from two adjacent and sometimes overlapping fields in robotics: developmental (or epigenetic) robotics (Asada et al. 2009) and HRI (Dautenhahn 2007). The studies were examples of developmental robotics in that many of the employed heuristics and the general experimental setup were informed by developmental psychology and psycholinguistics. The studies were largely human–robot interaction studies not only by virtue of using human–robot dyads as a basic experimental building block, but also by insisting on the use of naïve participants and unconstrained speech. To our knowledge, these experiments are, to date, the only studies in robotic language acquisition that utilize unconstrained speech as "learning substrate" for the machine-learning algorithms.

The rejection and prohibition experiments were jointly designed to test the hypotheses that the abovementioned *intent interpretations* (Ryan 1974) or paternal prohibitions (Spitz 1957), both described in Pea (1980), may form the developmental root of linguistic negation. The common denominator in both hypotheses is a requirement of affect, motivation, or volition. In the case of intent interpretations, caretakers—typically the parents—are thought to interpret bodily expressions of affect or volition linguistically, thereby providing the child with labels for its internal state, so to speak (Pea 1980). The mechanism behind Spitz' hypothesis is somewhat more complicated as it requires role reversal, but ultimately phrases accompanying (or embodying) parental forms or prohibition are thought to be the origin for the first negation words (Spitz 1957). Prohibition involves strong affect or motivation on both sides: the child must want some forbidden object or act, and act accordingly, whereas the parent must want to prevent the child from reaching or committing whatever is the object of its desire.

We, therefore, integrated a motivation module into an existing cognitive architecture and designed facial expressions and affectively congruent behaviors for positive, negative, and neutral motivation. The motivation would be triggered by randomly assigned object valences such that it appeared to the observer that the robot likes, dislikes, or "feels" neutral about the presented objects.

In terms of the experimental setup, participants were asked in all three experiments to act as language tutors for the child-like humanoid robot Deechee, an iCub robot (Metta et al. 2008). The majority of participants were recruited from the university campus and included members of staff and Ph.D. students from neighboring labs. None of the participants were roboticists and, in the negation experiment, all were naïve in the sense that they were not informed that the

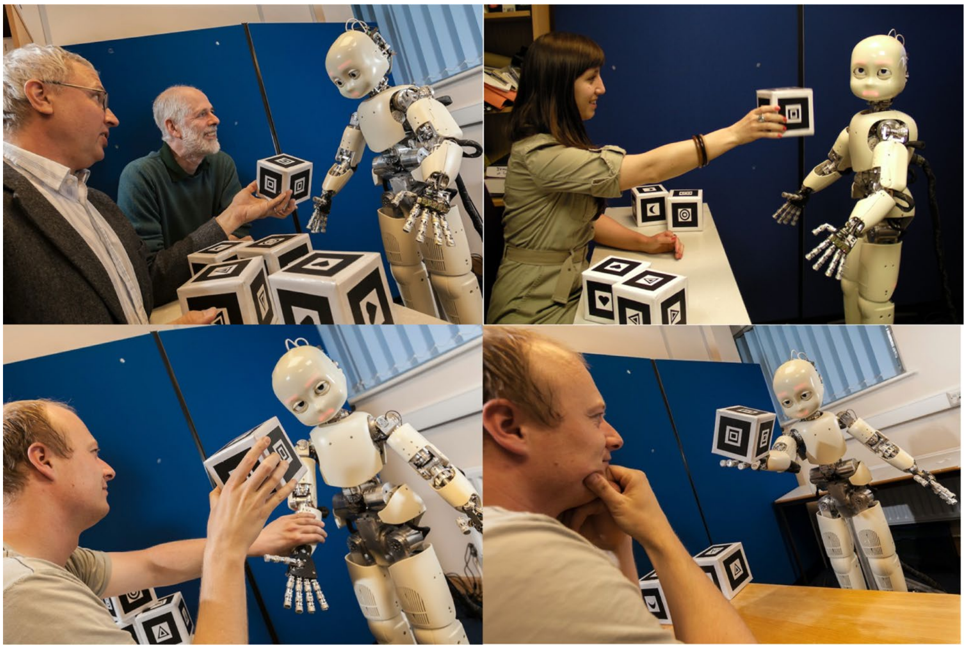

Fig. 1 Spatial layout and robot behaviors. Top left: watching or neutral behavior. Top right: rejective behavior triggered by objects with negative valence. Bottom left: physical restraint of the robots arm triggers negative affect. Bottom right: grasping/approaching behavior triggered by objects with positive valence. Top left and bottom photos are courtesy of Pete Stevens

topic under investigation was linguistic negation. Instead we adopted the participants instructions from Saunders' et al. (2012) experiment by telling them the more general story that they were acting as language tutors for the robot, and that they were ought to teach it the names of the shapes on the present boxes (cf. Fig. 1).

Deechee is an approximately 1.2-m-tall humanoid robot with 53 degrees-of-freedom and a range of sensors including visual (two cameras), auditory, haptic, proprioceptive, and torque sensors. Applications to control the robot are typically developed on top of the YARP middleware (Metta et al. 2006), and there exist a good number of software modules developed within a series of EU projects including modules to solve inverse kinematics, impedance control, amongst other, to allow for the design of robot behaviour at a relatively high level of abstraction.

In terms of the spatial layout, participants and robot were seated opposite each other with a table and labeled objects between them (cf. Fig. 1). Each participant was told that their role would be that of a language teacher for Deechee, and that they should teach it the names of the objects located between them. To increase the likelihood of participants engaging in a speech register akin to child-directed speech (CDS, Foulkes et al. 2005), we told them that they should imagine the robot to be a 2-year-old child. The lack of accuracy of speech recognition at the time required offline transcriptions of participants' speech. The prosody recognition, word extraction, and symbol grounding were performed offline as well. This meant that the teaching had to be broken up into sessions where the robot would learn, i.e. having its grounded vocabulary constructed based on participants' speech, in between these sessions.

Prosody recognition was used for the identification and extraction of prosodically salient words and involved the detection and measurement of pitch, energy, and duration for each word (see Saunders et al. 2011 for details).

Symbol grounding in the robotic context refers to the (technical) association or linking of symbols such as words or other linguistic units with non-symbolic data, typically data derived from the robot's own embodiment such as sensorimotor data or, in our case, data originating from a motivation-related subsystem (cf. Harnad 1990).

All three experiments consisted of five sessions per participant, but the session length in the negation experiments was approximately twice that of the session length in Saunders' experiment—5 and 2.5 min, respectively.

The differing session lengths necessitate a normalization of count data to allow for a comparison between the experiments. It is for this reason that we report normalized attribution counts and attributions per minute, in addition to the total counts in Sect. 2.

2.2 Lexical and semantic–pragmatic analyses

To extract all attributions of emotion, volition, and autonomy, the analysis proceeded according to the following steps:

Step 1: Merging of the three corpora into a single corpus and compilation of a global word list.

Step 2: Selection of words that could potentially be part of an attribution of emotion, volition or autonomy, and creation of a word list from this selection.

Table 1 Attribution words, word frequencies, and attribution categories

Word	Cat	#	Word	Cat	#	Word	Cat	#	Word	Cat	#
like	vol	902	boring	aut	1	indifferent	vol	2	glum	emo	1
want	vol	399	horrible	emo	4	nasty	vol	2	heart	aut	1
happy	emo	38	love	emo	4	play	aut	2	hungry	aut	1
sad	emo	31	smiley	emo	4	unhappy	emo	2	interest	vol	1
smile	emo	19	wake	aut	4	worry	emo	2	liking	vol	1
favourite	vol	15	enjoy	aut	3	ambivalent	vol	1	loved	emo	1
play	vol	15	enjoyed	emo	3	argue	aut	1	loved	vol	1
think	aut	10	funny	emo	3	argumentative	aut	1	miserable	emo	1
feel	aut	9	grumpy	emo	3	depressed	emo	1	mood	emo	1
keen	vol	9	interested	aut	3	despised	emo	1	nice	emo	1
hate	emo	8	love	vol	3	excited	emo	1	sorry	emo	1
interested	vol	7	tired	emo	3	fond	vol	1	teasing	aut	1
bored	emo	5	blood	aut	2	friendly	aut	1	terribly	vol	1
bored	vol	1	dislike	vol	2	frowny	emo	1	understand	aut	1
favorite	aut	5	feeling	emo	2	fun	emo	1	upset	aut	1
boring	emo	4	hurt	aut	2	fussy	aut	1			

The following words were extracted from a complete list of words from participants' transcripts of all three experiments. They formed the basis for extracting potential attributions of emotion (emo), volition (vol), or autonomy (aut) from the transcripts. The attribution categories were added post hoc, that is, once the utterances containing the stated words had been so classified

Step 3: Extraction of those utterances from the prosodic labeling files that contain at least one word of the list created in step 2.

Step 4: Categorical annotation or deletion of the utterances coming out of step 3. Deleted were those utterances that were not deemed to be attributions of one of the three types. All remaining utterances were categorized to be either attributions of emotion, attributions of volition, or attributions of autonomy.

Step 5: Categorization of utterances according to the target of attribution. Targets of attribution are either 'robot', 'participant' (=self), or 'interpersonal'.

Step 6: Selection of utterances whose attribution target is either 'robot' or 'interpersonal', and calculation of statistics based on this utterance set.

While steps 1, 3, and 6, should be self-explanatory, the steps 2, 4, and 5, need some more clarification.

In step 2, we selected all words that could potentially be part of one of the aforementioned types of attributions. Examples for emotion words are "sad", "happy", or "hate", examples for references to emotional expressions were "smile" or "frown". Typical words indicating the attribution of volition are "like" or "want". Somewhat more general is the list of words considered for attributions of autonomy: we included words that refer to the assumed presence of a metabolism ("hungry", "eat"), or allowing that a target's bodily functions could be impaired or its livelihood endangered ("hurt", "sick", "ill", etc.). In our view, the boundaries between the three categories—emotions, volition, autonomy—are not sharp, and there are several examples where a word could fall in either of these categories. Generally speaking, we tended to interpret attributions of emotion or volition as triggered by an indication in the here and now. Attributions of autonomy, on the other hand, could also refer to state of affairs in the past, the future, or be expressions of general preferences. Attributions involving the word "favourite", for example, were allotted to both the categories of attributions of volition as well as attributions of autonomy. With "favourite", the concrete choice of category hinged on our judgement whether the attribution was triggered by some behaviour on part of Deechee such as a frown or smile that just preceded the utterance, or whether it was a more general statement of preference that seemed more detached for the present situation. The outcome of the word selection is shown in Table 1.

For step 4, the same remarks hold as were just made for step 2. The categorization according to the three categories is certainly not sharp, so the precise numbers have to be regarded with a healthy amount of scepticism. However, we are in no doubt, that the selected utterances do make reference to at least one of the respective concepts and do express attributions of at least one the three kinds.

Step 5, the determination of the target of attribution is also not as trivial as it might seem. When participants use utterances such as "we don't like the squares today", it could be either an interpersonal attribution, expressing that the participant genuinely shares Deechee's sentiment about the square. However, it might also be some inverted form of the royal 'we' in which it is meant to refer to Deechee exclusively. In this case, the participant, fully aware of Deechee's

Table 2 Total and relative numbers of attributions of emotion, volition, and autonomy per participant for the rejection experiment (R1–R10), prohibition experiment (P1–P10), and baseline (= Saunders' experiment, B1–B9), accumulated across all five sessions

PID	# Attribution of emotion	apm	# Attributions of volition	apm	# Attributions of autonomy	apm
R1	0	0	3	0.12	0	0
R2	0	0	57	2.99	4	0.21
R3	14	0.63	96	4.30	1	0.04
R4	26	1.10	126	5.34	0	0
R5	12	0.45	146	5.43	0	0
R6	2	0.08	118	4.84	1	0.04
R7	21	0.84	95	3.79	3	0.12
R8	0	0	3	0.13	0	0
R9	5	0.20	44	1.74	4	0.16
R10	0	0	20	0.79	0	0
P1	1	0.04	34	1.33	0	0
P2	0	0	19	0.69	0	0
P3	5	0.20	150	5.93	2	0.08
P4	17	0.64	98	3.67	5	0.19
P5	1	0.04	52	2.06	0	0
P6	29	1.15	70	2.77	9	0.36
P7	8	0.31	50	1.94	3	0.12
P8	3	0.12	57	2.27	2	0.08
P9	1	0.04	60	2.28	0	0
P10	1	0.04	18	0.69	12	0.46
B1	0	0	0	0	0	0
B2	0	0	0	0	0	0
B3	0	0	0	0	0	0
B4	0	0	0	0	0	0
B5	0	0	0	0	0	0
B6	0	0	0	0	0	0
B7	0	0	0	0	0	0
B8	2	0.17	36	3.06	5	0.42
B9	0	0	11	1.06	0	0

Counted were only attributions to the robot and interpersonal attributions

PID participant ID, *apm* attributions per minute

severe limitations with regard to the capacity to express itself, may genuinely adopt the role of Deechee and speaks in its stead, so to speak.

3 Results

Section 3.1 lists and summarizes the tables with the absolute and relative counts of all those utterances produced by the participants that indicate that they did attribute volition, emotion, or agency more generally to Deechee, the robot. For the sake of brevity, we will in the following flatten the distinction between utterance and attribution and use the word 'attribution' for 'utterance that indicates attribution', and hope that this imprecision is transparent enough.

Section 3.2 lists three transcripts that provide some exemplary context in which such utterances were produced, and two additional transcripts that capture some of the more remarkable moments in the experiments. Transcript 4 captures aspect of a situation where a participant gets into an argument with Deechee because of its uttering "no", transcript 5 shows how even a speech-wise very disciplined participant who adhered to self-imposed restrictions in terms of his speech sometimes lapses in terms of these restrictions and makes reference to Deechee's preferences, indicating attribution of volition or intent.

3.1 Linguistic attributions of emotion, volition, and autonomy more generally within the three experiments

Table 2 lists the absolute and relative number of each attribution type and for each participant of the three experiments. As mentioned above, the important data for the purpose of

Table 3 Relative and total numbers of attributions of emotion, volition, and autonomy

	Rejection Exp	Prohibition Exp	Baseline
Emotion attributions	0.33 (80)	0.25 (66)	0.02 (2)
Volition attributions	2.95 (708)	2.35 (608)	0.50 (47)
Autonomy attributions	0.05 (13)	0.13 (33)	0.05 (5)
Total	3.34 (801)	2.73 (707)	0.57 (54)

Numbers are average attributions per minute (apm) across all participants of the respective experiment

Numbers in brackets: total sum of attributions of the respective category

comparison are the attributions per minute (*apm*), rather than the absolute number of attributions because the sessions of the negation experiments were approximately twice as long as the sessions in Saunders' et al. (2012) experiment. The measure of attributions per minute thus effectively factors duration out of account.

Table 3 lists the accumulated number of attributions, both apm and absolute numbers, for all three types and experiments.

As shown in Table 3, the average number of attributions per minute in the two negation experiments was considerably higher than in the baseline. In the rejection and prohibition experiments, the "average participant" produced nearly six and five times as many attributions of one of the three types as compared to the baseline, respectively. Attributions of volition outstrip attributions of emotion and attributions of agency by far in all three experiments. However, as Table 2 shows the accumulation of numbers across participants hides an important detail.

Rows B1–B9 of Table 2 clearly show that the production of relevant utterances was by no means evenly distributed across participants. Whereas participants B1–B7 produced no attributions of the relevant kind at all, B8 produced attributions of volition whose frequency was comparable to the average of the distribution in the negation experiments (cf. participants R10 or P10). Therefore, even though participants' speech was clearly impacted by Deechee's overt displays of emotion and volition, there is still considerable variability across participants. R1, for example, constituting the "lower limit" in terms of the number of attributions, only produced three such attributions across all five sessions, whereas participants such as R5 or P3 went "all out" and produced literally hundreds within a time frame of approximately 25 min without ever having prompted by the experimenters to do so.

The extremely skewed distribution in the baseline indicates that for most participants the attribution of emotion, volition, and autonomy seems to follow an "all or nothing" mechanism: either they do not produce any utterances of this kind, as is the case for B1–B7, or they produce a good number of them on par with participants where the robot displays emotions and preferences overtly. Participants such as R1 or R8, who produce only a very small number of such attributions are seen as exceptions.

3.2 Exemplary transcripts of attributions in context

To give a better impression of the conversational context in which the different attributions occurred, we provide several transcripts. Transcripts 1–3 are examples of the three types of attributions that we encountered during the experiments.

Transcript 1 illustrates the use of 'like', in this case negated, which we took to be an attribution of volition. It is the most frequently used word within this type of attribution (cf. Table 1), only followed by 'want'. From past research, we know that 'like' belongs to the prosodically most salient words linked to affect or motivation in the given experimental scenarios (cf Förster et al. 2019), and we see examples of this in the lines 5 and 11 of the transcript. As a consequence, 'like' frequently ended up in Deechee's active vocabulary and was produced by it in follow-up sessions.

Transcript 1: Attribution of volition. Participant R2, session 1, 25 s into the session. Prosodically salient words are in italic. Lines 5 and 11 show attributions of volition involving the word 'like'. Line 2 and 10 show negative intent interpretations expressed with 'no'. Capitalized words indicate the attribution of volition or motivation.

1. that *triangle*
2. *NO*
3. *what* about this
4. this *heart* sign
5. no, you don't *LIKE*
6. that *one*
7. what *about*
8. *squares*
9. *NOT* the squares
10. *NO*
11. you DON'T *LIKE* it
12. *put* that down

Transcript 2 illustrates the production of an attribution of a very strong emotion—hate (line 14). This was produced by participant R7 rather jokingly in this session, and R7 made produced lots of emotion attributions in this session, not merely "lexicalizing" Deechee's display of preferences but exaggerating them.

Transcript 2: Attribution of emotion. Participant R7, session 4, 4 min 37 s into session. Prosodically salient words are marked in italic, there are no prosodically marked words in lines 12–14, as the prosodic marking output is missing for these last few lines, probably as only the first 5 min of speech were considered for such marking, and these three *utterances are just beyond the 5-min mark. Lines 13 and 14 show strong attributions of emotion. Capitalized words indicate the attribution of motivation, volition, or emotion.*

1. yeah, the *target*
2. *oh*
3. you seem very *INDIFFERENT*
4. to it *actually*
5. *but* maybe INDIFFERENT
6. is too big a word for you at the *moment*
7. so, it's your *FAVORITE*
8. is this your *FAVORITE*
9. *yes,* your FAVORITE
10. *and*
11. *you're*
12. LEAST FAVORITE your
13. DESPISED you HATE the triangle
14. HATE HATE the triangle

Transcript 3 provides an example of the attribution of autonomy, the least frequent type of attribution. In line 4, participant P6 uses the word 'hurt' with respect to Deechee, indicating that it comes across as a being that could experience bodily harm. She could have used the word 'damage' here, which has a somewhat more mechanistic connotation, but chose the word 'hurt' instead. As mentioned in Sect. 0, and in line with more enactive interpretations of autonomy and agency (Brincker, 2016), we take references to bodily integrity and metabolism to be indicators of attribution of autonomy.

Transcript 3: Attribution of autonomy. Participant P6, session 2, 25 s into session. Prosodically salient words are marked in italic. 'Hurt' in line 6 was classified as attribution of autonomy as it makes reference to corporal health, the capacity to be hurt. From this line alone, one could not be sure whether it was indeed attributed to the robot or not, but on a later occasion in this session P6 uses 'hurt' again, and at that time uses 'you' as subject of that utterance. Capitalized words indicate attributions of motivation or attributions of autonomy more generally.

1. let me show you *another*
2. this *one*
3. this is the *triangle*
4. this shape this is quite a *sharp*
5. *shape*
6. could *HURT* you if you put your hands on all sides
7. it has three *equal* sides
8. this *triangle*
9. on this *one*
10. there are *three*
11. two *black*
12. one *white*
13. *this* one you're NOT VERY INTERESTED
14. IN are *you*

Transcripts 4 illustrates the "conversational power" of the combination of emotional displays with volition-indicating body behaviors which, in conjunction with the robot using the word 'no' itself, trigger an argument between participant R2 and Deechee. Deechee, which is "inaudible" in the transcript, keeps on saying 'circle' and 'no' repeatedly. Since 'no' was uttered just at the right time, R2 takes this as truth-functional denial, that is, that what he just said about the moon was wrong, and that it's actually called a circle. In lines 8 and 9, R2 gives a hint that he is aware of the ambiguity of the single-word utterance 'no': here, it could both mean truth-functional denial or, alternatively, be a case of motivation-dependent rejection of the object as a whole. R2, possibly due to the accidentally correct timing of the 'no' on part of Deechee, goes with interpretation 1 and berates Deechee for correcting him.

Transcript 4: "The argument". Participant R2, session 3, 3 min 59 s into session. Prosodically salient words are marked in italic. Some utterances are not marked as prosodically salient as the prosodic marking algorithm was slightly different with the very first participants and not every utterance "generated" a prosodically salient word. Capitalized words indicate attributions of motivation or volition, or attribution of autonomy more generally.

1. no, *ok*
2. yes, you're *Deechee*
3. ok, I put that
4. *down*
5. ahm, *and*
6. ok what *about*
7. the moon the crescent *there*
8. no, you DON'T LIKE that one
9. circles that's not a *circle* no, you DON'T LIKE it but it's not a circle
10. it's a crescent *that*
11. or a *moon*
12. if you'd *rather*
13. *no*
14. it's not a circle I can tell you *that*
15. it's not *circles*
16. DON'T ARGUE with me *Deechee*
17. ah that's a moon that *is*
18. ah no it's not *circles*
19. *no*
20. TRY TO BE ARGUMENTATIVE *Deechee*
21. *no*
22. it's not *circles*

Transcript 5 illustrates the difficulty of participants in refraining from making reference to the perceived intent or volition of the addressee. It also illustrates the conversational power of `no'. Participant P2 is a special case in that he changed his robot-directed speech dramatically between sessions 2 and 3. Via a personal contact, we were made aware that P2 had decided to optimize the learning progress of the robot by constraining his speech behavior in rather extreme ways. While he was still speaking more or less like other participants during the first two sessions, he only used a very small set of utterance types from session 3 onwards. By and large, he would then only provide object labels and give positive or negative feedback once the robot had uttered something, but not engage in any "chit chat" that, in his view, seemed to have no direct bearing on the learning task. As this violated our experimental instructions—to speak to the robot as if it was a 2-year-old child—we excluded his data in some strands of our analysis (cf. Förster et al., 2019). Nonetheless informative for the present analysis are those occasions where P2 slipped in terms of his self-imposed restrictions, and where he did engage in types of speech other than the mere provision of object labels or positive or negative feedback.

Transcript 5: "Loss of composure": R10, session 5, 1 min 20 s. Prosodically salient words are marked in italic. Violations of R10's self-imposed speech restrictions—labelling and feedback—can be seen in lines 5, 10, and 16. Capitalized words indicate attributions of volition.

1. *no*, don't say done
2. yeah, *triangle*
3. well *done*
4. no, I say *well* done
5. you *don't* say done
6. what's *this* one? Yes
7. a *heart*
8. yea
9. *crescent*, well done
10. stop saying *no*
11. circle
12. no square, yeah *square*
13. *well* done
14. *yeah*
15. *triangle*
16. do you *WANT* it? NO
17. YES
18. YEAH

While his slips in lines 5 and 10 could still be interpreted as being within his self-imposed restrictions, they are negatively reinforcing utterances about Deechee's speech after all, this is certainly not the case in line 16. Here, R10 felt compelled to make reference to Deechee's perceived volition. On several occasions, Deechee appeared undecided with respect to his preferences. Due to Deechee's object detection working less than perfectly well at times, Deechee's may have perceived the presented object as two objects between which his perception may have switched rapidly. If these two perceived objects happened to have opposite valences, approach and rejection behaviors would be triggered in rapid succession of each other. This could be interpreted as undecided behavior by a naïve observer. Yet, no one forced R10 to make reference to this behavior in his robot-directed speech, and such referencing seemed outside of his self-imposed limits if we compare it to the majority of his utterances. It is, therefore, interesting to observe that, if R10 showed any such lapses, they were typically lapses towards the attribution or referencing of volition.

4 Discussion

The results corroborate the anecdotal findings. If we assume that the participants express what they felt, they can be said to engage in a multitude of attributions of emotions, volition, and autonomy. The number of indications of such attributions in participants' speech in the negation experiments was four times more frequent than those in participants' speech from the "affect-less" baseline scenario. The concrete forms of attributions seem to be triggered in the here and now and thus appear to be identical to what Pea (1980), with reference to (Ryan 1974), termed *intent interpretations*. However, a deeper analysis involving a temporal alignment between the robot's emotional expressions and participants' speech is required to confirm the tight temporal coupling required for such a categorization.

While this may not seem surprising in view of the robot's overt affective or motivated behavior, it should be emphasized that these utterances are not mere detached and inconsequential comments referencing the robot's motivated behavior. Intent interpretations are tightly coupled with particular enactments of its behavior and appear to contribute to the determination of the next move in the interaction. The majority of attributions were attributions of volition that, following our categorization, are directly linked to or triggered by the robot's expressions and behavior.

But what role do these attributions or interpretations play in the conversation? Why do participants engage in them? While we cannot say for sure why humans engage in these attributions, and while the issue is, to our knowledge, undocumented in the developmental psychological literature, we offer a speculation.

We hypothesize that such “outspoken” attributions form part of a type of negotiation processes[3] whose purpose is the alignment of the two interactors’ wills and emotions, and the joint determination of the next move in the interaction (cf. Cowley 2005). Within our simple learning scenario, the next move will nearly always consist of the choice of the next object to talk about. Participants, by overtly and audibly attributing the robot’s intent, provide it with a near-optimal slot for “having its say”, or for slotting in its contribution.

Imagine a situation where Deechee would have the capacity to express its “feelings” or attitudes in a more verbose manner. Participants’ attributions, especially attributions of volition, would afford Deechee with an occasion to “jump in” and correct the participant’s interpretation of its own will, if Deechee wanted to change the course of action. It appears that those participants that engage in them, by offering up their interpretations of the robot’s intent audibly, provide the robot with a conversational scaffold to take its turn in the joint negotiation, once it has the capacity to do so (see, for example, transcript 4, lines 6–9, or transcript 5, lines 16–18).

It is unclear to us as to whether the tight temporal coupling of intent interpretations with the corresponding motivation-indicating body behaviors is a feature specific to conversationally asymmetric speakers such as parent and child, or tutor and robot, or whether it can be more widely observed. Although underplayed in accounts on language acquisition centered around joint attentional frames, the tight temporal relationship between what is attentionally picked out and its “linguistic label” appears to be of some importance. In interactive machine learning, focusing on language or speech such couplings have to the best of our knowledge been ignored in the past.

The above findings of audible attributions of autonomy in their various forms are in line with our findings with respect to negation words, particularly ‘no’. In addition, there, the production of negative utterances skyrocketed once the robot displayed emotion, volition, or intent. Many of the negative utterances could be interpreted as attributions of volition akin to the ones we found in the current study. The main difference here is that negation words such as ‘no’ typically do not have standard semantic meaning that could be easily mapped to emotion, volition, or autonomy more generally. However, on the pragmatic level, many of the utterances appear to fulfill the same purpose—to guide the joint action of the dyad and to align each other’s preferences. Whether a participant utters “that’s a sad face” and puts the object down, or whether she says “no, you don’t like it” and puts it down, does pragmatically not really make a difference, even though the semantics of ‘no’ and ‘sad’ appear to differ considerably.

In terms of language acquisition, and following Pea’s suggestion (Pea 1980), our study provides further evidence for the hypothesis that intent interpretations may lie at the root of children’s first negation words as well as being the root of the first words for expressing emotion or volition. Both negation and emotion or motivation-related words such as “like” are often prosodically salient in the tutor’s speech, as we have already documented in Förster et al. (2019). Under our operationalization in the robotic architecture, this means that these words are extracted, grounded in sensorimotor-motivational data, and propagated into the robots active vocabulary.

Notably, many of these words also belong to the earliest vocabularies of human toddlers (cf. Fenson et al. 1994).

Nevertheless, a deeper analysis is required to determine and quantify the precise temporal relationships between the robot’s bodily expressions and participant’s linguistic interpretation or attribution thereof, as issues of timing are likely to have a considerable impact in terms of the robot’s learning success.

5 Future work

The present analysis is only the first step in a planned series of analyses of emotional and volitional robot-directed speech as collected during the two negation experiments. Akin to the multi-layered analysis that we performed in the work reported there, we would like to analyze the presented attributive or interpretive utterances more deeply around a conversation analytical transcription of parts of the interaction. Further, an independent part will consist of a temporal analysis investigating the precise temporal alignment between the robot’s emotional and behavioral expression and participants’ attributive utterances.

Independent of the work on language acquisition is the question in how far emotional and volition-indicating expressions could or should be used to synchronize human–robot joint actions in more applied scenarios such as object handovers which the authors are currently working on.

A major difference between our language teaching or tutoring setup and more applied scenarios where humans do not act as teacher is the presence or absence of some form of interactional symmetry. Teaching or tutoring setups are characterized by a marked difference in the

[3] Using the term ‘negotiation’, we do not mean to imply this to be a conscious decision on part of the human or a construal of the robot as entity that is genuinely capable of engaging in full-blown negotiations. A good part of the observed patterned and interpersonal behavior might be largely subconscious. Therefore, the “decision” to offer the robot a slot for “having its say” may not be a conscious decision at all but rather a form of ingrained interaction pattern, on the level of conversational turns. If this is the case, it may take more conscious effort to not engage in this type of behavior than just to “go with the flow”.

relevant skill set between human and robot, and the human teacher seems to take a role of both higher responsibility and also of higher power. In this sense, the two interactants are not equals in such setups and there is a marked asymmetry between them. It is unclear whether humans in more symmetrical human–robot setups would engage in linguistic attributions of autonomy to the same degree, as those in our experiments, thereby offering the robot a slot to "have its say".

6 Conclusions

Originating from a set of human–robot interaction experiments, we have documented major differences in participants' speech which were apparently trigged by the robot's emotional expressions and congruent body behaviors. The relevant changes in their speech indicated that these participants attributed emotion, volition, or autonomy more generally to the robot as compared to a "non-affective" baseline scenario. However, we also found that 2 out of 9 participants in the baseline scenario produced similar utterances indicating that such attributions were not strictly dependent on an overtly "affective stance" on part of the machine.

Numerically, the three attribution types were dominated by attributions of volition, the majority of which involved the words 'like' or 'want'.

We hypothesize that those forms of attribution that are triggered by behaviors in the here and now, namely attributions of emotion and volition in our classification, are by and large identical to what Joanna Ryan referred to as *intent interpretations*. Being unaware of any research into the function of these utterances, we hypothesize that they are meant to afford the linguistically less competent learner a scaffold to contribute and clarify their stance in the negotiation of an ongoing joint action. It is unclear whether such scaffolding is restricted to competence-wise asymmetrical dyads, where the more competent interactant offers the less competent interactant a hand, or whether we also find such conversational scaffolds in the negotiation processes of more symmetrical setups.

Our observation hints towards the possibility that such attributions may be necessary in certain forms of human–robot joints action. It is, however, not clear how such "low-level" conversational attributions relate to the "deeper" unidirectional attributions of emotions towards robots that have been mentioned by Scheutz (2011). While the latter could pose a danger for our collective psyche by robots distracting humans from healthy human–human relations and monopolizing their users' attention and care, the former might be necessary to get certain forms of joint actions done.

Funding The experiments from which the corpora under analysis in this paper originate were supported by the EU Integrated Project ITALK ("Integration and Transfer of Action and Language in Robots") funded by the European Commission under contract number FP-7-214668. The analysis presented in the current paper was supported by The Turing Institute and within the project 'Learning collaboration affordances for intuitive human–robot interaction'. The Turing provides its partners with funding for activities that enhance the research development more generally among data and AI researchers, enhance the Turing Fellow community and enhances existing Turing Fellowship projects.

Data availability The corpus of transcripts originating from the rejection experiment has been made publicly available here: https://doi.org/10.18745/DS.18196. Some exemplary videos of the human–robot interaction in the negation experiments can be found here: https://doi.org/10.18745/pb.21637.

Compliance with ethical standards

Conflict of interest The authors do not declare any conflicts of interest.

References

Asada M et al (2009) cognitive developmental robotics: a survey. IEEE Trans Auton Ment Dev 1(1):12–34

Briggs G, Scheutz M (2014) How robots can affect human behavior: investigating the effects of robotic displays of protest and distress. Int J Soc Robot 6(3):343–355

Brincker M (2016) Dynamics of perceptible agency: the case of social robots. Mind Mach 26(4):441–466

Buss S (2002) Personal autonomy. The Standford Encyclopedia of Philosophy

Carlson Z et al (2019) Perceived mistreatment and emotional capability following aggressive treatment of robots and computers. Int J Soc Robot 11:727–739

Cowley SJ (2005) Languaging: how babies and bonobos lock on to human modes of life. Int J Comput Cogn 3(1):44–55

Dautenhahn K (2007) Socially intelligent robots: dimensions of human-robot interaction. Philos Trans R Soc B 362(1480):679–704

Epley N, Akalis S, Waytz A, Cacioppo JT (2008) Creating social connection through inferential reproduction: loneliness and perceived agency in gadgets, gods, and greyhounds. Psychol Sci 19(2):114–120

Fenson L et al (1994) Variability in early communicative development. Monogr Soc Res Child Dev 59(5):i–185

Ferrari F, Paladino MP, Jetten J (2016) Blurring human-machine distinctions: anthropomorphic appearance in social robots as a threat to human distinctiveness. Int J Soc Robot 8(2):287–302

Fischer K (2011) Interpersonal variation in understanding robots as social actors. s.l., IEEE, pp. 53–60

Fischer K, Foth K, Rohlfing K, Wrede B (2011) Mindful tutors: linguistic choice and action demonstration in speech to infants and a simulated robot. Int Stud 12(1):134–161

Fischer K, Lohan K, Foth K (2012) Levels of embodiment: linguistic analyses of factors influencing HRI. Boston, IEEE, pp. 463-470

Förster F (2018) Coding scheme for negative utterances. University of Hertfordshire, Hatfield

Förster F, Saunders J, Nehaniv CL (2018) Robots that say "no" Affective symbol grounding and the case of intent interpretations. IEEE Trans Cogn Dev Syst 10(3):530–544

Förster F, Saunders J, Lehmann H, Nehaniv CL (2019) Robots learning to say "no": prohibition and rejective mechanisms in acquisition of linguistic negation. ACM Trans Hum Robot Interact 8(4):26

Foulkes P, Docherty G, Watt D (2005) Phonological variation in child-directed speech. Language 81(1):177–206

Gahrn-Andersen R (2020) Seeming autonomy, technology and the uncanny valley. AI & Soc. https://doi.org/10.1007/s00146-020-01040-9

Geiskkovitch DY, Cormier D, Seo SH, Young JE (2016) Please continue, we need more data: an exploration of obedience to robots. J Hum Robot Interact 5(1):82–99

Harbers M, Peeters MM, Neerincx MA (2017) Perceived autonomy of robots: effects of appearance and context. A World with Robots, pp. 19–33

Harnad S (1990) The symbol grounding problem. Phys D 42(1–3):335–346

Hoffmann L, Bock N, Rosenthal vd Pütten AM (2018) The peculiarities of robot embodiment (EmCorp-Scale): development, validation and initial test of the embodiment and corporeality of artificial agents scale. Chicago, USA, Association for Computing Machinery, pp. 370–378

Jochum E, Millar P, Nuñez D (2017) Sequence and chance: design and control methods for entertainment robots. Robot Auton Syst 87:372–380

Kamide H, Kawabe K, Shigemi S, Arai T (2013) Development of a psychological scale for general impressions of humanoid. Adv Robot 27(1):3–17

Levinson SC (1995) Interactional biases in human thinking. In: Goody EN (ed) Social intelligence and interaction. Cambridge University Press, Cambridge, pp 221–260

Levinson SC (2006) On the human "interaction engine." roots of human sociality: culture, cognition, and interaction. Berg Publishers, Oxford, pp 39–69

Metta G, Fitzpatrick P, Natale L (2006) YARP: yet another robot platform. Int J Adv Rob Syst 3(1):8

Metta G et al (2008) The iCub humanoid robot: an open platform for research in embodied cognition. Gaithersburg, USA

Papenmeier F, Uhrig M, Kirsch A (2019) Human understanding of robot motion: the role of velocity and orientation. Int J Soc Robot 11(1):75–88

Pea RD (1980) The development of negation in early child language. The social foundations of language & thought: essays in honor of jerome bruner. W. W. Norton, New York, pp 156–186

Rietveld E, Kiverstein J (2014) A rich landscape of affordances. Ecol Psychol 26(4):325–352

Rosenthal-von der Pütten AM, Bock N, Brockmann K (2017) Not Your Cup of Tea? How Interacting With a Robot Can Increase Perceived Self-efficacy in HRI and Evaluation. Vienna, IEEE, pp. 483–492

Ryan J (1974) Early language development: towards a communicational analysis. In: The integration of a child into a social world. Cambridge University Press, London

Saunders J, Lehmann H, Sato Y, Nehaniv CL (2011) Towards Using Prosody to Scaffold Lexical Meaning in Robots. Frankfurt am Main, IEEE, pp. 1–7

Saunders J, Lehmann H, Förster F, Nehaniv CL (2012) Robot acquisition of lexical meaning—moving towards the two-word stage. San Diego, IEEE, pp. 1–7

Schaefer KE, Foots AN, Straub ER (2018) Applied robotics for installations and base operations: user perceptions of a driverless vehicle at fort bragg. Technical Report: ARL-TR-8265. US Army Research Laboratory, Aberdeen Proving Ground, Maryland

Scheutz M (2011) The inherent dangers of unidirectional emotional bonds between humans and social robots. In: Lin P, Abney K, Bekey GA (eds) Robot ethics: the ethical and social implications of robots. MIT Press, pp 205–221

Spitz RA (1957) No and yes: on the genesis of human communication. International Universities Press, New York

Tomasello M (2000) The social-pragmatic theory of word learning. Pragmatics 10(4):401–413

Tomasello M (2003) Constructing a language: a usage-based theory of language acquisition. Harvard University Press, Cambridge

Vanman EJ, Kappas A (2019) "Danger, will robinson!" the challenges of social robots for intergroup relations. Soc Pers Psychol Compass 13(8):e12489

Varela FG, Maturana HR, Uribe R (1974) Autopoiesis: the organization of living systems, its characterization and a model. Biosystems 5(4):187–196