"I See What You Did There": Understanding People's Social Perception of a Robot and Its Predictability

BOB R. SCHADENBERG, DENNIS REIDSMA, and DIRK K. J. HEYLEN, University of Twente, the Netherlands

VANESSA EVERS, University of Twente, the Netherlands and Nanyang Technological University, Singapore

28

Unpredictability in robot behaviour can cause difficulties in interacting with robots. However, for social interactions with robots, a degree of unpredictability in robot behaviour may be desirable for facilitating engagement and increasing the attribution of mental states to the robot. To generate a better conceptual understanding of predictability, we looked at two facets of predictability, namely, the ability to predict robot actions and the association of predictability as an attribute of the robot. We carried out a video human-robot interaction study where we manipulated whether participants could either see the cause of a robot's responsive action or could not see this, because there was no cause, or because we obstructed the visual cues. Our results indicate that when the cause of the robot's responsive actions was not visible, participants rated the robot as more unpredictable and less competent, compared to when it was visible. The relationship between seeing the cause of the responsive actions and the attribution of competence was partially mediated by the attribution of unpredictability to the robot. We argue that the effects of unpredictability may be mitigated when the robot identifies when a person may not be aware of what the robot wants to respond to and uses additional actions to make its response predictable.

CCS Concepts: • **Human-centered computing** → **Interaction design theory, concepts and paradigms**; • **Computer systems organization** → **Robotics**;

Additional Key Words and Phrases: Predictability, responsive actions, human-robot interaction, social perception

ACM Reference format:

Bob R. Schadenberg, Dennis Reidsma, Dirk K. J. Heylen, and Vanessa Evers. 2021. "I See What You Did There": Understanding People's Social Perception of a Robot and Its Predictability. *Trans. Hum.-Robot Interact.* 10, 3, Article 28 (July 2021), 28 pages.
<https://doi.org/10.1145/3461534>

1 INTRODUCTION

Predictability in human-robot interactions is considered to be essential for humans to understand robots [85], for coordinating actions with robots [86, 87], for improving task performance [52],

This work has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 688835 (DE-ENIGMA).

Authors' addresses: B. R. Schadenberg, D. Reidsma, and D. K. J. Heylen, University of Twente, Drienerlolaan 5, 7522 NB Enschede, the Netherlands; emails: {b.r.schadenberg, d.reidsma, d.k.j.heylen}@utwente.nl; V. Evers, University of Twente, Drienerlolaan 5, 7522 NB Enschede, the Netherlands; email: v.evers@utwente.nl.

This work is licensed under a Creative Commons Attribution International 4.0 License.

© 2021 Copyright held by the owner/author(s).

2573-9522/2021/07-ART28 \$15.00

<https://doi.org/10.1145/3461534>

safety [1, 45], and trust in the robot [43, 58], and can support decision-making processes for human-in-the-loop robotic systems [69]. While predictability is an important concept to human-robot interactions, there is no common understanding about the concept of predictability and there are various ways to approach this concept—it is an umbrella concept. Predictability is used in terms of it being a *property of robotic motion* [4, 23, 24, 55, 61, 93]. In this context, the predictability of a motion relates to how well it matches the person’s prediction regarding the robot’s motion trajectory, given that the goal of the robot is known [23]. Studies also differ in *what about the robot’s behaviour* a person is predicting, where a robot can be predictable in one aspect, but unpredictable in another. For instance, Dragan et al. [23] differentiate between “predictable motions,” which are optimised to allow the prediction of the robot’s motion trajectory, and “legible motions,” which are optimised to allow the prediction of the goal the robot is trying to achieve with its motion. While these motions can result in the same motion in certain situations, in highly ambiguous situations, legible motions allow a person to predict its goal, but its trajectory is unpredictable, and vice versa for predictable motions [24]. Following on from predicting the end-state of a robot action, other studies are about predicting what action the robot will take after the on-going action [28, 61, 65, 95, 102]. And finally, people can also *attribute predictability* as a quality of a robot [23, 26, 28, 30, 31, 65]. These different conceptualisations of predictability generate different hypotheses, operationalisations and design recommendations, limiting us to effectively take predictability into account in the development of robots.

Given the importance of predictability of robot behaviour, little research has been done on how to actually design for predictability in human-robot interactions. Current literature on designing for predictability mostly relates to the predictability of the robot’s motions. However, to achieve predictable human-robot *interactions*, we need to move beyond generating single robot actions that are predictable, and look at how such actions relate to the robot’s previous actions and to the environment in which the interaction takes place. Unpredictability in interactions can stem from different sources and may involve different design solutions. While certain robot actions may be an obvious choice given the robot’s internal state, they may catch a person interacting with the robot off-guard. The robot could see things that are invisible to the human eye, calculate countless potential reasons to explain its perception, and finally compute the most optimal action given the situation. This is very different from human cognition, making it difficult for people to try and understand a robot by using their theory of mind—a set of cognitive mechanisms that allow people to predict the mental states of others [75], which are also used to make sense of robot behaviour [22, 83, 107]. Critically, even when a robot facilitates people in predicting its future actions, this does not necessarily mean that this robot is also considered to be more predictable by people [28]. It is not well understood to which extent the attribution of predictability and the ability to predict robot behaviour each contribute to the **Human-Robot Interaction (HRI)** experience.

In social HRI studies, the effect of seemingly unpredictable robot actions—as described above—on people’s social perception of the robot show mixed results. For social interactions with robots, some studies report that a degree of unpredictability in robot behaviour can be desirable in facilitating engagement and increasing the attribution of mental states to the robot [57, 79, 81, 89]. However, the unpredictable actions have also been reported to negatively affect people’s social perception of the robot in that it was judged to be less reliable, technically competent, understandable, and trustworthy [80], or show less patience [65]. In part, the mixed results in whether a degree of unpredictability is desirable or not may be explained by different operationalisations of predictability. As a result, the unpredicted actions have different consequences for the person interacting with the robot, influencing the person’s social perception of the robot. Individual differences can also play a role in the different effects of robot predictability in that people are not all equal in their desire for predictability [11]. For people who have a higher desire for predictability—such as

autistic children [6, 15, 66]—the effects of unpredictable robot actions may have stronger negative effects.

All things considered, our current conceptual understanding of robot predictability—as it relates to HRI—cannot explain how the predictability of a robot affects people's social perception of a robot, nor how each aspect of predictability is related to this. This limits us in effectively taking robot predictability into account in the design of robot behaviour. In this article, we report on two studies. First, we report on the development of a new measurement tool for measuring to what extent someone associates unpredictability with a robot. This tool was also used in the main study—an online video-HRI study. In the main study, we investigate how a robot's predictability relates to the observers' social perception of a robot. To study which aspect of robot predictability matters more to the observers' social perception of the robot, we look at both the observers' attribution of predictability as a quality of the robot, as well as their actual capability to predict the robot's future action. Subsequent to investigating the relationship between robot predictability and observers' social perception of the robot, we investigated to what extent the observers' social perception of the robot is influenced by their intolerance of uncertainty, and by the knowledge that their perception of the robot's surroundings is incomplete.

The remainder of this article is structured as follows. First, we will elaborate on how predictions play a role in human perception, what these insights mean for predictability in HRI, and on the individual differences. This literature forms the basis upon which we develop our research questions and hypotheses, described in Section 3. In Section 4, we then report on our measurement tool for the attribution of predictability to a robot, which is used in the main study. The method of the main study is described in Section 5, which is followed by our analysis of the robot's predictability, as well as the relation between the robot's predictability and participants' social perception of the robot, and their intolerance of uncertainty in Section 6. We conclude the article with a discussion in Section 7 on how we interpret the results of our study and answer our research questions. We conclude what this means for our understanding of robot predictability and how we design robot behaviour in Section 8.

2 BACKGROUND

2.1 Predictability and Robots

2.1.1 Predictability (and Related Concepts). Depending on the perspective that is taken, various terms are being used to describe (the various stages in) how people understand robot behaviour. This process itself is often referred to as explainability or interpretability [64]. In the field of Artificial Intelligence, this process is described in terms of legibility [9, 27], predictability [52], explanation [22, 92], transparency [33], readability [95], and anticipation [40]. In this article, we look at the explainability of robots from a cognitive science point of view, as in the field of cognitive science, there already exists many decades worth of research on how people understand the world around them. Insights from this field may also help us to interpret how people understand robots in specific [64]. In cognitive science, perception is generally viewed as a problem of causal inference [e.g., References 3, 21, 41, 46]. That is, the sensory information received by the brain is caused by objects and processes in the environment. Based on this sensory information, the brain needs to infer what caused it. Predictions, and thus the predictability of objects and processes in the environment, are believed to play a central role in how the brain solves the problem of causal inference (predictive processing, see References [16, 38, 39, 47, 56, 77]). This is based on the premise that the brain continually generates predictions on what input comes next based on current input and learned associations [41, 46]. When describing how people understand robots, we will therefore talk in terms of predictability. In the remainder of this section, we will provide a conceptualisation

of predictability and how it relates to people predicting robot behaviour, based on insights from these cognitive theories.¹

Predictability is not a static property of a robot and is dependent on the experience of the person interacting with it. When people interact with a certain robot (or robot type) for the first time, their initial predictions might not be very accurate as they have no prior experience with this robot upon which to base the predictions. As a result, how people eventually perceive the robot is less influenced by their expectations, and instead relies more strongly on the sensory information coming from the robot (top-down versus bottom-up processing) [32, 105]. Thus, a robot is inherently less predictable when interacting with it for the first time, but people's predictions also influence their perception to a lesser degree. The predictability of a robot may therefore be less impactful for understanding the robot during such initial interactions than for longer-term interactions.

As people interact with a certain robot, their predictions will be continuously evaluated in terms of their accuracy (through perceptual inference [77] or active inference [8, 37, 39]). When predictions are incorrect, the internal models in the brain that generate these predictions are updated accordingly to generate more accurate predictions in the future [71, 106]. Whether the predictions will become more accurate through this process of learning depends, in part, on the extent in which the robot's past behaviours are indicative of its future behaviour. In other words, the need to be structural regularities in the robot's behaviour. When people are able to discern these regularities, they can use them to update the internal models. The predictability of a robot therefore relates to how easily people can perceive the structural regularities in its behaviour so that they may learn to accurately predict its behaviour—predictability relates to facilitating this learning process. Note that this process of evaluating predictions shows similarities with the concept of explanation (i.e., how people explain robot behaviour).

As we alluded to in the introduction, predictions differ in what aspect of the robot's behaviour they are about (e.g., actions, motions, intentions). The internal models that generate predictions are believed to be utilised throughout the brain—from low- to high-level perceptual processing [101]. This implies that people also generate predictions on different aspects of the robot's behaviour [50], or in different levels of detail [53]. Critically, lower-level predictions (e.g., predictions regarding robot's motion trajectories or sounds) influence higher-level predictions (predictions regarding actions or intentions) and vice versa [101]. Thus, when people may predict the robot to take a certain action (a higher-level prediction), they simultaneously may also predict certain robot motions required for this action.

Taken together, the predictability of a robot shapes how people perceive and interact with it. When interacting with a robot for the first time, it will be less predictable, but our understanding of the robot also relies less strongly on its predictability. As people interact with the robot, they can learn to predict its behaviour, and the predictions will start to play a larger role in our understanding of the robot. The predictions relate to all aspects of the robot—from low-level motion primitives to high-level beliefs, desires, or intentions. Following from the cognitive science perspective on how people understand robots, we define *predictability* as: “the ability to quickly and accurately learn to predict the behaviour of a robot.” To what extent a robot is *predictable* depends on (1) what degree the robot's behaviour shows structural regularities, and (2) how easily these structural regularities can be discerned by people. What the predictions will be about can differ in abstraction. They can be very concrete when they relate to what motions, sounds, or smells to expect, but might also be more abstract in that the prediction is only about what action the robot will take, or what it intends to do, and not how.

¹For a more detailed description of how we applied these cognitive theories, please see Chapter 6 in Schadenberg [84].

2.1.2 Behavioural Predictability and Attributed Predictability. We might expect that more predictable robots are also *perceived* as such by the people interacting with them. While this may often be the case, it may not always be so. To clearly distinguish between the predictability of a robot and the perception thereof by people, we will refer to these two concepts as, respectively, *behavioural predictability* and *attributed predictability*. Results from Driggs-Campbell and Bajcsy [28] illustrate the difference between attributed and behavioural predictability. In their study, participants were asked to predict when an autonomous car would change lane, where the car's motions were either mechanical/robot-like, or human-inspired. Even though participants were slower to predict the lane-changing behaviour of the car with human-inspired motions, they considered this car to be more predictable than the car with mechanical/robot-like motions. Critically, the results from this study imply that optimising for behavioural predictability may lead to different designs from optimising for attributed predictability.

A key question is how the concepts of behavioural and attributed predictability relate to each other. In general, people are good at dealing with unpredictability, thus some aspects of behavioural predictability (e.g., lower-level predictions on motions versus higher-level predictions on actions or intents) may be more important for the attributed predictability. The context in which a robot performs an unexpected action may also matter for the relationship between behavioural and attributed predictability. For instance, when a person is very familiar with a robot, any unexpected actions will have a larger impact than when a person interact with the robot for the first time. In the latter case, the person does not know what to expect from the robot from prior experiences. This lack of information on which to base predictions results in inaccurate predictions, as well as a high degree of uncertainty in the predictions [16, 37, 39, 47]. Due to this uncertainty, the predictions will have a reduced influence on perception, and *less attention is given any unexpected sensory information* [32]. Considering that these predictive processes are also believed to play a role in people's social perception [7, 96], this weighting of the impact of predictions in proportion to their uncertainty on attention may similarly influence the relationship between behavioural and attributed predictability. That is, when predictions regarding the robot's behaviour are known to be uncertain, people pay less attention to any unexpected actions, which may reduce their impact on (social) perception (e.g., attributed predictability) is reduced.

2.2 Effects of Predictability on Human-robot Interactions

While robots can be designed to facilitate predictability, a degree of unpredictability will be unavoidable in many cases. A robot with high structural regularities in its behaviour—making it easier to learn to predict its behaviour—can be programmed by making it do the same thing, in exactly the same manner, again and again. However, for robots that need to interact with people, the robot will likely need to do more to perform its function, increasing its complexity and making it more difficult to perceive the structural regularities in its behaviour. Robot behaviours such as those resulting from responding to people or the environment, where the responsive action can appear out of thin air. This occurs when the event the robot is responding to is interpreted differently by the person interacting with it, when this person is not aware of the event, or when the event is perceived faultily. For some robot applications, unpredicted actions will have clear (severe) negative effects, such as leading to dangerous situations, or reducing trust in the robot. For such cases, it is then vital to improve the robot's behavioural predictability. However, for robots used for social interaction, results reported in literature are not as clear cut. The remainder of this article will be about robots used for social interaction, thus we limit this section to discussing the effects of behavioural unpredictability for the people interacting with the robot.

With robots used for social interaction, a *degree* of behavioural unpredictability may have positive effects. Salem et al. [79] asked their participants to help a robot with unpacking boxes. The

robot assisted the participant by providing information on where to put the items, where it showed either unimodal, congruent multimodal, or incongruent multimodal behaviour. The results indicated that despite decreasing task performance, the robot that showed incongruent multimodal behaviour was preferred by participants, in that it was considered more humanlike, more likeable, and the participants experienced greater shared reality with the robot. The authors argue that the incongruency of the robot's multimodal behaviour led to a degree of (behavioural) unpredictability, which was explained by participants attributing mental states to the robot, such as that the robot was "cheeky" or tried to fool them.

In a similar vein, robots that sometimes behave unpredictably have been found to increase behavioural engagement of adults [89] and children [57, 81]. In the study conducted by Short et al. [89], participants had to interact with a robot and play a rock-paper-scissors game. At a certain point in the interaction, the robot would verbally conclude that it had won the game, while the rules of the game dictated that it had lost. It did so three out of twenty times the game was played, making it inconsistent in its behaviour and therefore more difficult to predict its behaviour. The results indicated that while the inconsistent robot was judged as being less fair than when it showed consistent behaviour, it also increased behavioural engagement.

The positive effects of behavioural unpredictability are explained by arguing that it makes robots more humanlike, which in turn increases the attribution of mental states to the robot [79]. Intuitively this explanation makes sense, as humans are unpredictable in their behaviour. Highly predictable behaviour may also be considered as boring, whereas unpredicted actions could spark interest [19]. However, also in social interaction, unpredictable robot behaviour can lead to attribute negative attributes to the robot. Salem et al. [80] compared a non-erroneous robot with a robot that showed erratic behaviour. The latter was operationalised by navigating in an erratic manner and by playing a song that was not selected by the participant. Participants considered the erratic robot to be less reliable, less technically competent, less understandable, and less trustworthy. Interestingly, the erratic robot was not perceived as more humanlike in this study.

One issue with studies on predictability and robots is that predictability is operationalised very differently. As a result, the unpredicted actions have different consequences for the person interacting with the robot. For example, the person may not be able to smoothly carry out their task, because the robot does not comply with his/her commands [57, 65], is making mistakes [57, 65], or is illegible in its actions [79]. In case of robots showing erratic navigation, while unpredictable, this is also not behaviour that would naturally occur in humans. This might explain why these behaviours were not perceived as being more human-like, which is argued to be the reason why a degree of unpredictability may lead to positive effects. Therefore, concluding that it is the robot's *predictability* that influences people's social perception of the robot, seems premature.

While the studies above made the robot differ in its behavioural predictability, this was not the main purpose of the studies—it was not assessed whether the behavioural predictability was successfully manipulated, nor was the attributed predictability assessed. It is also unclear to what extent the behavioural and attributed predictability led to the reported findings.

2.3 Individual Differences in Generating and Using Predictions

One factor that may influence how the predictability of a robot shapes people's perception of it are individual factors. Generating predictions is ultimately an individual process, influenced by social factors such as the person's desires, affective states and stereotypes [70]. Thus, what is predictable to one person, may be unpredictable to another. The mechanism for generating and using predictions also differs per person, resulting in differences between people in their ability to deal

with unpredictability. For instance, deviations in the prediction mechanism in the brain have been linked to explaining autism spectrum conditions [54, 72, 90, 99], delusions and hallucination in schizophrenia [18, 35] and depression [49]. A person's ability to deal with unpredictability is also a key component in explaining anxiety and is linked to various anxiety pathologies [see Reference 10, for a review]. In this field, unpredictability is often referred to as uncertainty. Uncertainty and unpredictability are highly similar constructs and are often used interchangeably [42]. Uncertainty is often used to reflect a person's internal state, whereas unpredictability is more quantitative and amenable to experimental manipulation, and is often used to describe the nature of a stimulus. **Intolerance of uncertainty (IU)** is defined as "an individual's dispositional incapacity to endure the aversive response triggered by the perceived absence of salient, key, or sufficient information, and sustained by the associated perception of uncertainty" [10], and comprises two components, namely, a "desire of predictability" and "uncertainty paralysis" [5]. The ability to predict the future allows people to anticipate and increase the odds of desired outcomes, while avoiding or bracing themselves for future adversity. When faced with uncertainty, this ability is diminished and reduces how efficiently and effectively one can prepare for the future. In turn, this can lead to anxiety, and in more extreme cases to an anxiety disorder.

While people, in general, differ in their intolerance of uncertainty [11], autistic individuals score much higher on their intolerance of uncertainty than the general population [6, 15, 66]. This is in line with current theories on autism spectrum conditions, which place their ineffectiveness or inability to deal with predictions at the core of the condition, resulting in them having difficulty dealing with unpredictable (volatile) environments [54, 72, 90, 99]. For robot applications aimed at supporting autistic individuals, the predictable nature of robots is often argued to be one of the main strengths of robots for these individuals [e.g., References 19, 20, 29, 48, 82, 97]. For individuals who are less well equipped to deal with unpredictability, such as autistic individuals, it may be especially important to design robots to be predictable.

3 PROBLEM STATEMENT

3.1 Main Study

3.1.1 Goal of the Main Study. The aim of our main study is to investigate how the visibility of the cause of a robot's responsive actions influences the following: behavioural predictability, attributed unpredictability,² an observer's social perception of the robot and whether this is mediated by the behavioural and attributed predictability, and how an observer's IU influences the observer's perception of the robot. For our methodology, we opted for a video-HRI study with a third person perspective. According to the predictive coding framework, people constantly generate predictions to make sense of sensory information. Thus this also applies to watching a video of a human-robot interaction. Moreover, the video-HRI methodology has been used extensively to study predictability and legibility [e.g., References 24–26, 59, 68, 95].

3.1.2 Predictability Operationalisation and Manipulation. In our study, participants watched a video of an interaction between a person and a robot unfold. While the participants will not be able to predict the robot's behaviour initially, the interaction follows a highly structured format. The structural regularities in the robot's behaviour should therefore be relatively easy to discern. For participants who are able to do so, this would presumably improve the accuracy of their predictions related to the robot's behaviour. However, for some participants, depending on the condition they are in, the robot will perform actions that cannot be predicted. These unexpected actions might

²For the remainder of this article, we refer to the attributed unpredictability, rather than attributed predictability, as this is what we specifically measured.

make it more difficult to discern the structural regularities in the robot's behaviour, resulting in less accurate predictions.

The unexpected actions are operationalised through responsive actions. These could either be predicted based on the situational context, or could not be predicted because of a lack thereof. The situational context was manipulated by whether the participants could see what event the robot was responding to (the visibility of the cause; our independent variable). The events—the cause of the robot's responsive action—were either visible, obstructed from view, or there were no events. We consider this as a natural and realistic way to introduce unpredictability in the robot's behaviour, as in real interactions with a robot it may not always be apparent to what a robot is responding to.

When the cause is visible, participants can explain and may even predict the responsive actions. In turn, this should make it easier to derive that the responsive actions are an exception (i.e., they hold no predictive value for future behaviour) and make it easier to discern the structural regularities in the robot's behaviour. When the cause is not visible, either because the event was obstructed or there was no event visible, the robot's action cannot be predicted, nor is there evidence regarding an explanation as to why the robot performed this action. This should make it more difficult to derive the structural regularities in the robot's behaviour. In case of the cause being obstructed from view, the participant has perceptual information that the information of the scene is incomplete, and it is therefore ambiguous whether the robot is responding to an event or not. It is therefore made explicit that any predictions are based on limited information. This might increase the uncertainty of any predictions and thereby reduce the impact of predictions on the (social) perception of the robot. For this condition, we are interested in understanding how knowing that you cannot fully understand the reasons behind the robot's behaviour influences a participants' social perception of the robot and the attributed predictability.

3.1.3 Research Questions. The first set of research questions relates to the robot's predictability in relation to the visibility of the cause of responsive robot actions. Inspired by the findings of Driggs-Campbell and Bajcsy [28], we are interested in how unpredictable robot actions influence the behavioural and attributed unpredictability depending on whether the participant see a cause for the robot's unexpected behaviour.

Topic 1: The first questions concern the relation between the *visibility of the cause of responsive robot actions* and the *robot's predictability*.

RQ 1a To what extent does the visibility of the cause of responsive robot actions influence the *behavioural predictability* of the robot?

RQ 1b To what extent does the visibility of the cause of responsive robot actions influence the *attributed unpredictability* of the robot?

We assume that when the cause of the responsive actions is not visible, the actions are unpredictable and make it more difficult to assess the structural regularities in the robot's behaviour. That is, the unpredictable actions do not have predictive value for predicting future actions, and hence should be discarded. We therefore expect that participants are less accurate and less confident in their prediction when the event was obstructed or not there, compared to when the event is visible—the behavioural predictability is reduced. Our hypotheses related to the attributed unpredictability are based on the notion of weighting the impact of predictions according to their uncertainty [16, 37, 39, 47] and how this directs people's attention [32] (see Section 2.1.1). We expect that the participants will associate unpredictability with the robot more strongly when the cause of the responsive actions is not visible (because the event is not there or is obstructed),

compared to when it is. Presumably, uncertainty regarding predictions reduces their impact on (social) perception. We therefore expect that in the ambiguous condition, where participants know they do not have full perceptual information of the scene, such ambiguity will mitigate their attribution of unpredictability. Regarding the relation between behavioural predictability and attributed unpredictability, we do expect that they are correlated with each other.

Our second set of research questions relates to the participants' social perception of the robot. Social perception was measured in terms of warmth, competence, and discomfort. While discomfort is a dimension unique to robots [13], warmth and competence are considered universal dimensions of human social cognition and are used to judge whether another person is a threat or a friend [34]. Together, these dimensions are believed to explain how people characterise others [104]. This leads to our second research question and sub-research questions:

Topic 2: The second set of research questions concerns *the impact of the visibility of the cause of responsive robot actions on the participants' social perception of the robot, and the extent to which this is mediated by the robot's predictability*

RQ 2a To what extent does the visibility of the cause of responsive robot actions influence a participant's social perception of the robot (i.e., what social attributes are attributed to the robot)?

RQ 2b To what extent does the robot's *behavioural predictability* mediate the social perception of the robot?

RQ 2c To what extent does the *attributed unpredictability* mediate the social perception of the robot?

We expect that when the cause of the responsive actions is not visible, the robot will negatively influence the participants' social perception of the robot, compared to when the cause is visible. Moreover, we expect that when participants know that they do not have full perceptual information, this will mitigate a reduction in warmth and competence and an increase in discomfort. Again, these hypotheses are based on the notion of weighting the impact of predictions according to their uncertainty [16, 37, 39, 47] and how this directs people's attention [32]. Based on the results from Driggs-Campbell and Bajcsy [28], we expect that the participants' social perception of the robot is mediated more strongly by the participants' attributed unpredictability of the robot. Given that we expect that behavioural predictability and attributed unpredictability are correlated, we also expect such a mediation effect for behavioural predictability, although one that is less strong than for attributed unpredictability.

The last of our research questions relate to a possible moderation effect of the participants' IU, given the relation of this construct with predictability [42] and populations that are known to have difficulty dealing with unpredictability, such as autistic children [6, 15, 66]. This leads to our third research question and sub-research questions:

Topic 3: The final set of research questions concern the impact of the *participants' intolerance of uncertainty* on the (possible) *relations between the visibility of the cause of responsive robot actions and the participants' social perception of the robot or the robot's predictability*.

RQ 3a To what extent does the participants' intolerance of uncertainty moderate the *attributed unpredictability* of the robot?

RQ 3b To what extent does the participants' intolerance of uncertainty moderate the participants' social perception of the robot?

Given that those higher in IU respond more negatively to instances where they have insufficient information [10], we expect that they will also respond more strongly to the robot's responsive

actions when the cause is not visible in terms of their social perception of the robot and attributed unpredictability.

3.2 Measuring Attributed (Un)Predictability

Previous studies measured attributed predictability either through a single item [65], or multiple items specific to predictability of robot motion [23, 26]. Single-items are more vulnerable to random measurement errors and unknown biases in the meaning and interpretation of that item. With multiple-item scales, the random measurement error is more likely to be cancelled out. Moreover, they cover a broader range of meanings of a construct, which can reduce the effect of differently interpreting an item. We therefore developed a new multi-item scale that measures to what extent unpredictability is attributed to a robot, and which is not restricted to the predictability of robot motions. The next section describes the method and results of this development.

4 DEVELOPMENT OF THE ATTRIBUTED UNPREDICTABILITY SCALE

To achieve our aims, we need to be able to measure the extent to which a person associates the attribute “(un)predictability” with the robot. In this section, we describe how we came to our measurement method.

4.1 Materials and Methods

To populate our multi-item scale of attributes that may indicate a robot’s predictability, we first looked at English synonyms, antonyms, and related words for “predictable” and “unpredictable” using Merriam-Webster’s thesaurus [63], resulting in 22 items. We then asked participants to rate these according to how well they fit their intuitive concept of predictability. We analysed their answers to arrive at a reduced set of items to be used in our main study.

4.1.1 Participants. Participants were recruited through Amazon’s Mechanical Turk (mTurk), which resulted in 99 adults (42 men, 57 women) participating. Three male participants were excluded from the analysis, because they indicated that they did not know what “predictable” or “unpredictable” means. Participants received \$0.50 for participating. To meet the recruitment criteria, participants had to be from the United States of America, have completed 500 or more assignments on mTurk, and have an approval rate of 99% or higher on mTurk.

4.1.2 Procedure. We informed participants that we are developing a measurement tool for measuring the “predictability” of a robot. To that end, we wanted to know to what extent they associated certain words with the general concept of a robot that is “predictable” and a robot that is “unpredictable.” For items related to “predictable,” participants were asked “Using the scale provided, how closely are the words below associated with a robot that is predictable.” The same question was used for items related to “unpredictable,” but with the different phrasing emphasising that the robot is unpredictable. The provided scale was a 7-point Likert scale, from 1 = *definitely not associated* to 7 = *definitely associated*. In case the participant did not know the word, they could choose *I do not know what this word means* instead. The items were presented in a randomised order.

We intentionally did not provide any description or visualisation of robots, as the questionnaire should serve as a general measure of the attributed predictability of robots. Moreover, we did not provide any definition of predictability, taking it as a type of *projective latent content* where it is assumed that most people understand, and to a certain degree share, a common meaning of the construct and one wants to approach that intuitive meaning rather than limit the user to restrictive definitions [74].

4.2 Results

Prior to performing the analysis, we removed 4 items (mutable, immutable, mercurial, and capricious) for which more than 5% of the participants answered that they did not know what the words meant. We then conducted a **Principle Component Analysis (PCA)** on 18 items with varimax rotation. This resulted in four components, of which two were related to the items for predictability, and the other two were related to unpredictability. Given that unpredictability is the inverse of predictability, we expected one component on which both predictability and unpredictability items would load. Instead, it is likely that they loaded on different components because of asking both positively and negatively worded questions. This can produce an artefactual two component structure in the absence of two constructs [62, 78, 91]. Therefore, we proceeded with only the items related to unpredictability, for which we had more items than for predictability.

We performed a PCA on 10 unpredictability items (see Table 1) with varimax rotation. The KMO measure verified the sampling adequacy for the analysis (KMO = 0.87), and for the individual items the KMO values were > 0.82. Bartlett's test of sphericity indicated that correlations between items were sufficiently large for PCA ($\chi^2(45) = 577.90, p < 0.001$). Two components had eigenvalues above Kaiser's criterion of 1. These two components explained 68.67% of the variance. Correspondingly, the scree plot (see Figure 1) shows a point of inflexion at the third component. Table 1 shows the mean scores on how closely each item is associated with an unpredictable robot, and the component loadings for each item, where we highlighted items with loadings of 0.6 or higher and with a cross loading difference greater than 0.2. Component one seems to represent unpredictability associated with robots, as the mean association scores are high for these items. Component two contains items that had relatively low association scores. It may therefore be indicative of a type of unpredictability that is less suited to ascribe to a robot, and was discarded from our set of items.

4.3 Final Multi-item Scale for Attributed Unpredictability

In our main study, reported in the remainder of this article, we used the resulting measurement tool and proceeded with measuring *attributed unpredictability* through the six items of component one (unpredictable, irregular, inconsistent, random, variable, and erratic), taking the average of those items as the score for attributed unpredictability.³ Note that, as this is a dictionary study, conducted with participants from the United States, these six words may have different meanings to people from other countries. When using this new scale in other countries, we need to be mindful of this.

5 MAIN STUDY: MATERIALS AND METHODS

5.1 Participants

Participants were recruited through Amazon's Mechanical Turk platform. In total, 169 adults (74 men, 94 women, 1 non-binary) participated through mTurk. They received \$1.20 for participating. To meet the recruitment criteria, participants had to be from the United States of America, have completed 100 or more assignments on mTurk, have an approval rate of 99% or higher on mTurk, and not have participated in our study for the questionnaire development reported in the previous section. Participants were asked how well they could hear the robot on a 7-point Likert scale, from 1 = *I didn't understand anything the robot said* to 7 = *I understood every word the robot said*. Overall, participants could understand the robot's speech well ($M = 6.48, SD = 0.70$). Eight participants who rated the robot's speech with a 4 or lower were excluded from the analysis, because it was essential for this study that they could clearly hear what the robot was saying. Another four participants

³Also, see the supplementary files for the attributed unpredictability scale.

Fig. 1. Scree plot of the PCA on the unpredictability items.

Table 1. The Component Loadings for the Items on Unpredictability and the Mean Scores for How Closely Each Items is Related with an Unpredictable Robot

Item	<i>M (SD)</i>	Component 1	Component 2
Unpredictable	6.55 (1.27)	.98	−.24
Irregular	6.09 (1.36)	.85	.04
Inconsistent	6.35 (1.13)	.84	−.06
Random	6.35 (1.11)	.78	.05
Variable	6.29 (0.96)	.65	.12
Erratic	6.28 (1.14)	.64	.35
Unreliable	5.98 (1.37)	.50	.41
Inconstant	5.93 (1.37)	.47	.53
Skittish	4.32 (1.87)	−.07	.87
Fickle	4.32 (1.74)	.02	.84
Eigenvalue		5.63	1.24
% of variance		56.31	12.38
Cronbach’s α		.90	.72

were excluded from the analysis, because their completion time was too fast to realistically assume they watched the video until the end. Their answers also corroborated this assumption. We also asked participants to what extent they had prior experience with robots, and included this factor as a covariate in our analysis. However, this factor did not influence any of the outcome variables, and thus we included participants regardless of their prior experience with robots. As a result, the analyses reported in this article are based on the responses of 153 adults (62 men, 90 women, 1 non-binary).

This study was reviewed and approved by the ethics committee of the faculty of Electrical Engineering, Mathematics, and Computer Science of the University of Twente, and is registered under reference number “RP 2018-52.” Prior to participating in the study, participants first had to sign an informed consent form.

5.2 Materials

For this study, we used Robokind's R25 humanoid robot called "Zeno," or "Milo." The robot's speech is generated by the integrated Acapela Text-To-Speech engine and uses the American English children's voice "Josh," and includes lip synchronisation. The interaction was recorded with a 1080p camera and lasted 1 minute and 15 seconds. To improve the audio quality, we recorded it separately for the robot and for the adult using two microphones, and combined the audio feeds with the video feed afterwards.

5.3 Interaction Design

The participants watched a short interaction of a person interacting with a humanoid robot. This interaction was the same for *all* participants. In the video, the robot is playing a storytelling game with an adult, where the robot tells a story in which the adult is the main character. The game proceeds in a very structured manner: the robot tells something, asks input from the adult, and then uses his answer to proceed with the story in an obvious direction. This basic pattern is repeated 3 times, and it forms the main structural regularity in the interaction that can be used to predict the robot's actions. The participants viewing the video received no information telling them what the interaction was about to prevent priming them to pay too close attention to regularities in the interaction.

The interaction pattern (what happens in the video) can be seen in Table 2. The interaction consists of an introduction, the storytelling game, and two events to which the robot responds. At the start of the video, the robot introduces itself and asks whether the person would like to start. The robot then starts telling the story, which is about the adult entering a kitchen and finding his favourite food. The story is divided into three parts that each follow the same fixed interaction pattern. First, the story proceeds. Second, the robot asks the adult a question related to the subject of the previous sentence about the story. Finally, the robot uses the adult's answer to resolve what happens with the adult and the subject. The video ends prior to the robot's response to the adult's answer to the third question ("*Are you hungry?*").

Twice during the interaction, the robot would display a responsive action to an event (the cause) that occurred within the situational context. The events are (a) the adult receives a text message on his cellphone and starts reading it, and (b) another person walks in and distracts the adult from the interaction with the robot.

5.4 Experimental Design

The study was set up as a 3 by 1 between-participants design. The independent variable in our study is the visibility of the cause of the robot's responsive actions (which we will reference to as *visibility of the cause* for brevity), which had three levels. We measured four dependent variables, namely, behavioural predictability, attributed predictability, robot social attributes, and the intolerance to uncertainty.

Screenshots of the video for each condition can be seen in Figure 2. In one condition, the events were visible, and therefore the responsive actions could be predicted. We refer to this condition as the *visible cause* condition. In the other two conditions, the events were not visible. This was either because there were no events, or because we limited the view of the participant shortly after the start of the interaction. We refer to the former as the *no cause* condition. In this condition, the responsive actions cannot be predicted. Moreover, the adult in the video who is interacting with the robot shows no response to the robot's responsive behaviour. This could indicate to the participant that there is simply no event that explains the robot's responsive actions. In the condition where we limited the view of the participant, the responsive actions are not predictable,

Table 2. The Interaction Pattern of the Storytelling Game

INTRODUCTION			
Introduction	“Hi, my name is Zeno.” (Robot waves) “Shall we start?” [Adult]“Okay.” *Mask is applied smoothly*		
STORYTELLING GAME			
	<i>Story part 1</i>	<i>Story part 2</i>	<i>Story part 3</i>
Story introduction	“You stand in front of a small door.”	“There is a distinct smell in this room.”	“When you look around, you see a pizza lying on the table.”
	First event here		
Story question	“How tall are you?”	“What is your favourite food?”	“Are you hungry?”
	Second event here		
Adult’s answer	[Adult] “Six foot.”	[Adult] “Pizza.”	[Adult] “Yes, I am.”
Story resolution	“You open the door, but have to crouch to get through it. You enter a kitchen.”	“The smell reminds you of what a pizza smells like.”	*Mask is removed, revealing the adult* and *Video ends*
EVENTS AND CORRESPONDING RESPONSIVE ACTIONS			
First event	[Adult receives text message and views it on his mobile phone.] “Could you please put that away.” [Adult complies.] “Thank you.”		
Second event	[Another person taps the adult on shoulder. The robot turns head to the new person.] “Please go away, we are busy here.” The other person leaves. The robot turns head back to the adult.] “Thank you. Let’s continue. What is your favourite food?”		

but participants know that they do not have full perceptual information of the interaction. We limited the view by applying a mask that obstructed the view of the adult interacting with the robot. The mask was applied smoothly just after the introduction. This way, the participant knew that there was a person interacting with the robot. After the mask was applied, participants would still hear the adult reply to the questions of the robot. We refer to this condition as *ambiguous*, as the robot may or may not be responding to actual events. Note that the interaction followed the exact same script for each condition with the same two events. We only manipulated the perceptual information available to participants.

5.5 Procedure

For the study itself, participants were instructed that they would watch a video of a person interacting with a robot and that afterwards they would need to answer a couple of questions regarding the content of the video, their impression of the robot, and some background questions. Prior to viewing the video, participants saw a test video where the robot would speak for 2 min, so that they could adjust their volume settings if this was required to hear the robot’s speech clearly. Next, participants viewed one of three videos (one for each condition), which was selected randomly. After viewing the video, we first asked the participants to write down the last thing the robot said. This was followed by questions on behavioural predictability, how well the participants understood what the robot was saying, the **Robotic Social Attribute Scale (RoSAS [13])**, our new attributed

(a) Visible cause: perceptual information regarding the cause of the robot's responsive actions are visible.

(b) Ambiguous: no perceptual information regarding the cause of the robot's responsive actions is visible because the event is obstructed by a mask.

(c) No cause: no perceptual information regarding the cause of the robot's responsive actions is visible. Instead, the adult provides perceptual information that there is no cause as he does not respond to the responsive actions.

Fig. 2. The three conditions of the study, which differ in whether the participant can see the cause of the robot's responsive actions. In this case, another person distracting the adult from the interaction with the robot.

unpredictability scale from the previous section, and finally some background questions. The order in which the items of the RoSAS and attributed predictability questionnaires appeared were randomised between participants.

5.6 Measurements

5.6.1 Behavioural Predictability. Similar to Takayama et al. [95], we used an open question format to measure behavioural predictability, and asked how confident participants were in their prediction on a 7-point Likert scale. The open question ("What do you expect the robot will say next?") was coded in terms of the presence of a keyword related to "eating" and to "pizza." Answering this question requires the participant to connect the robot's question ("whether the adult is hungry") to the latest story development ("a pizza lying on the table"). When none of the keywords were present, the question was marked with a 0. When one keyword was present, related to either eating or pizza, behavioural predictability was marked with a 1, or with a 2 when keywords relating to both eating or pizza were present.

5.6.2 Attributed Unpredictability. To measure to what extent participants associate unpredictability with the robot, we used the self report questionnaire with the six unpredictability items described in Section 4.

Fig. 3. Raincloud plots (scatter, box, and density plot) that show the participants' confidence in their prediction (behavioural predictability) for each of the three conditions.

Fig. 4. Raincloud plots (scatter, box, and density plot) that show the scores on the attributed unpredictability of the robot for the three conditions.

5.6.3 Robot Social Attributes. Participants' social perception of the robot was measured through the RoSAS [13]. This questionnaire measures the robot's warmth, competence, and caused discomfort.

5.6.4 Intolerance of Uncertainty. The participants' intolerance of uncertainty was measured through the **Intolerance of Uncertainty Scale—Short form (IUS-12 [12])**. This is a 12-item questionnaire, for which we report the mean of the items.

6 RESULTS

6.1 Behavioural Predictability

There was a ceiling effect on the accuracy of the predictions, as most participants were able to make an accurate ($n = 127$), or partially accurate prediction ($n = 19$), and only 11 participants did not make a correct prediction. As a result, we used Fisher's exact test to account for the three cells that had values less than 5, and found no significant difference between the accuracy of the prediction and the visibility of the cause ($p = 0.16$).

Figure 3 shows the participants' confidence in their prediction on what the robot would say next. A one-way ANOVA indicated no significant difference between the participants' confidence in their prediction for the three conditions ($F(2, 154) = 0.18, p = 0.83$). The average confidence scores when the cause was visible, ambiguous, or no cause condition were, respectively, 3.93 ($SD = 1.60$), 4.10 ($SD = 1.60$), and 3.94 ($SD = 1.65$). Given that there is no statistically significant relation between the visibility of the cause and behavioural predictability, there is also no mediation effect of behavioural predictability on the relation between visibility of the cause and the participants' social perception of the robot.

6.2 Attributed Unpredictability

The reliability analysis of the attributed unpredictability questionnaire indicated that the internal consistency was excellent (Cronbach's $\alpha = 0.90$). The scores for the attributed unpredictability of the robot under each of the three conditions can be seen in Figure 4. Levene's test indicated that the variances for attributed unpredictability were significantly different for the cause visible condition ($F(2, 150) = 11.37, p < 0.001$). As the assumption of homogeneity of variance was not

met, we used Welch's adjusted F ratio and the Games-Howell post hoc procedure. There was a significant effect of the visibility of the cause on the attributed unpredictability as determined by one-way ANOVA (Welch's $F(2, 90.65) = 22.20, p < 0.001, \hat{\omega}^2 = 0.22$). The participant's **intolerance of uncertainty (IU)** did not appear to moderate the relation between visibility of the cause and attributed unpredictability, as adding an interaction effect resulted in a non-significant result ($F(2, 152) = 0.14, p = 0.874$). Post hoc comparisons revealed that participants rated the robot as being less unpredictable when the cause was visible ($M = 12.65, SD = 5.54$) compared to the ambiguous condition ($M = 19.45, SD = 10.01, p < 0.001, d = 0.86$), or when no cause was visible ($M = 20.53, SD = 7.40, p < 0.001, d = 1.21$). There was no significant difference between the ambiguous and no cause conditions ($p = 0.769$).

6.3 Robotic Social Attributes

We expected that the participants' IU may influence their social perception of the robot. We therefore included IU as a covariate in the analysis of the three RoSAS subscales, given that the IU was not significantly related to the conditions, as determined by one-way ANOVA ($F(2, 150) = 0.49, p = 0.612$). IU was measured using the IUS-12 questionnaire, which had an excellent internal consistency (Cronbach's $\alpha = 0.90$). The assumption of homogeneity of variance was met for all three of the one-way ANCOVA's on the RoSAS subscales reported below. The assumption of homogeneity of regression slopes was also met, as IU showed no significant interaction effects with visibility of the cause. To assess whether IU moderated the relationship between visibility of the cause and the participants' social perception of the robot, we also ran the ANCOVA's with an interaction effect. The RoSAS subscales all had high internal consistency scores (Cronbach's α for warmth = 0.85, competence = 0.90, discomfort = 0.79). The mean score and confidence intervals of the RoSAS subscales can be seen in Figure 5.

6.3.1 Warmth. We found no significant effect of the visibility of the cause on warmth ($F(2, 149) = 0.22, p = .801$). Additionally, IU was not significantly related to warmth ($F(1, 149) = 0.24, p = 0.625$), nor did it interact with the visibility of the cause ($F(2, 149) = 2.04, p = 0.134$).

6.3.2 Competence. For competence, there was a significant difference between the visibility of the cause ($F(2, 149) = 7.78, p = 0.001, \omega^2 = 0.08$) after controlling for IU. Note that although IU was neither significantly related to competence ($F(1, 149) = 1.24, p = 0.267$) nor did it interact with the visibility of the cause ($F(2, 149) = 0.97, p = 0.381$), we nevertheless kept it as a covariate. Post hoc comparisons with a Bonferroni correction revealed that the robot was considered to be more competent when cause was visible ($M = 6.80, SD = 1.48$), compared to when it was ambiguous ($M = 5.76, SD = 1.62, p = 0.003, d = 0.67$), or when there was no cause ($M = 5.78, SD = 1.51, p = 0.003, d = 0.68$). The difference between the ambiguous and no cause condition was not significant ($p = 1.000$).

As there was a significant difference in competence attributed to the robot under the condition of the cause being visible, we proceeded with a mediation analysis to investigate whether this difference is mediated by attributed unpredictability. Note that we cannot carry out a mediation analysis with behavioural predictability, because we did not find a significant difference between visibility of the cause and behavioural predictability (through quality or confidence). The mediation analysis was conducted in SPSS using the PROCESS macro [44]. To interpret the b coefficients, note that for visibility of the cause, we coded the condition "cause visible" as 0, "ambiguous" as 1, and "no cause" as 2. Competence and attributed unpredictability are both continuous variables with a value between 1 and 9.

(a) Warmth subscale of the RoSAS, showing to what extent participants associated warmth with the robot.

(b) Competence subscale of the RoSAS, showing to what extent participants associated competence with the robot.

(c) Discomfort subscale of the RoSAS, showing to what extent participants associated discomfort with the robot.

Fig. 5. Raincloud plots (scatter, box, and density plot) that show the scores on the three RoSAS subscales for each condition.

The direct effect of visibility of the cause on competence was significant ($b = -0.32$, $t(149) = 2.04$, $p = 0.044$). To assess whether there can be a mediation effect (an indirect effect) of attributed unpredictability between the visibility of the cause and competence, the b coefficient between visibility of the cause and attributed unpredictability needs to be significant, as well as the b coefficient for attributed unpredictability and competence. This is the case, as the regression of visibility of the cause on attributed unpredictability was significant ($b = 0.65$, $t(150) = 5.11$, $p < 0.001$), as well as the regression of attributed unpredictability and competence ($b = -0.29$, $t(149) = -3.02$, $p = 0.003$). To calculate the indirect effect, we used bootstrapping with 5000 samples to generate the confidence interval for the b coefficient, which was -0.19 (95% CI $[-0.35, -0.05]$). This is significant at an α of 0.05 as the confidence interval does not include 0. As both the direct and indirect effect of visibility of the cause on competence are significant, attributed unpredictability only partially mediates this relationship. The mediation model with the regression coefficients can be seen in Figure 6. For all these steps, we controlled for IU. However, this variable was not significant in any of the steps.

6.3.3 Discomfort. We found no significant effect of the visibility of the cause on discomfort ($F(2, 149) = 1.94$, $p = 0.147$). However, IU is significantly related to discomfort ($F(1, 149) = 17.15$, $p < 0.001$, $\omega^2 = 0.10$), but it did not interact with the visibility of the cause ($F(2, 149) = 1.51$, $p = 0.224$). Participants with higher scores on IU associated more discomfort with the robot.

Fig. 6. Mediation model for predicting competence based on visibility of the cause with attributed unpredictability as mediator. The model includes the b coefficients and standard error for each relationship.

7 DISCUSSION

In this study, we investigated the relationship between the visibility of the cause of a robot's responsive actions and action predictability, attributed predictability, the social attributes associated with the robot and the participants' intolerance of uncertainty. Our results indicate that not seeing the cause of the robot's responsive actions results in a stronger association of unpredictability with the robot, than when the cause was visible. We find no evidence that ambiguity in whether the robot responded to an actual event mitigated the effect of not seeing to what external event the robot responded. Interestingly, the participants' attributed unpredictability mediated the relation between the visibility of the cause and the attribution of competence to the robot. That is, participants rated the robot as more unpredictable, which also made them judge the robot as less competent. Considering a robot as competent is important, as it has been shown that robots that are considered less competent are also considered less trustworthy [36]. The mediation effect of attributed unpredictability was a partial mediation effect, as not seeing the cause of the robot's responsive actions also directly decreased the participants' attribution of competence to the robot. Even so, this direct effect was much smaller than the mediation effect of the attributed predictability. A possible explanation of the direct effect of seeing the cause of the robot's responsive actions is that they may have been perceived as mistakes. In turn, this could further decrease the robot's competence, as mistakes have been found to decrease the competence of a robot [76]. Unlike Dragan and Srinivasa [26], who found that higher familiarisation (leading to higher behavioural predictability) leads to more comfort, we found no effects between the visibility of the cause, attributed unpredictability and discomfort, nor with warmth.

We found no evidence that knowing that you cannot fully understand the reasons behind the robot's behaviour (in the ambiguous condition) mitigated the effect of not seeing the cause of the responsive robot actions on the attribution of competence. We hypothesised that this would mitigate the effect of unpredictable responsive actions, due to a reduced influence of predictions on social perception given the increased uncertainty regarding the predictions. Possibly, the explicit knowledge that limited information was available upon which to base predictions did not further decrease the already uncertain predictions due to seeing the robot for the first time. Alternatively, this may have to do with the fact that this was a video study in which participants did not interact with the robot. Prior research found that people try to make sense of a robot more strongly when they expect to interact with the robot in the future [30, 31]. Possibly, participants in our study were

less inclined to make sense of the robot's behaviour as they did not interact with it nor would they in the future.

For behavioural predictability, we expected that not seeing the cause of the robot's responsive actions would make it more difficult for participants to infer the structural regularities of the interaction, and therefore make it more difficult to predict the robot's actions. We found no evidence that participants had difficulty predicting the robot's behaviour, as most participants were able to accurately predict the robot's next action. Neither did the visibility of the cause influence the participants' confidence in their prediction. This means that after watching just over one minute of interaction, they were able to discard the unpredictable actions as noise and use the structural regularities in the interaction to generate an accurate prediction. Unlike in the study of Takayama et al. [95], our participants did not know what questions they would need to answer prior to watching the interaction, and were only given the instruction that they would view an adult interacting with a robot after which they would need to answer a couple of questions. Combined with a brief interaction, participants may not have had enough information upon which to make a reasoned prediction and went with their "gut feeling."

Reflecting back on how we manipulated the robot's behavioural predictability in our study, we wonder whether the manipulation was strong enough to find a difference, even though our pilot indicated that it was. The goal of our manipulation was to make the robot differ in its behavioural predictability in a realistic interaction, to limit confounding factors such as behaviour that is clearly erroneous or detrimental to the task at hand. Thus, we opted for responsive robot actions, which are actions that are at risk of being perceived as unpredictable when the event the robot is responding to is not clear—a scenario that may well occur in real HRI's. In previous studies on predictability in HRI, behavioural predictability was also measured by asking participants to make a prediction [28, 60, 61, 95, 102] and indicate their certainty of their prediction [95]. However, in these studies, the robot's behavioural predictability was manipulated through limiting the information available to participants by having the robot perform fewer communicative actions [95] or by presenting different and less information in a graphical user interface [102], or through how the robot moved [28, 60, 61, 102]. Unlike these studies, we manipulated the structural regularities in a conversation, where the participants had to predict robot speech. The downside of this is that conversations need to be fairly structured to still make sense, limiting us in how unpredictable we could make the interaction whilst keeping it realistic. In that sense, manipulating robot behaviour through robot motions may be better suited for manipulating behavioural predictability without making the scenario unrealistic or unnatural.

Regarding the individual differences between people in their intolerance of uncertainty, we had expected that this would moderate the effects of the visibility of the cause and the participants' social perception of the robot. We hypothesised that those high in IU would respond more strongly to not seeing the cause of the robot's responsive actions, which would negatively influence their social perception of the robot. We found no evidence that this is the case. IU only influenced the discomfort associated with the robot. Robots are a novel technology and are likely to induce uncertainty, which in turn may cause more discomfort for those who have a higher intolerance of uncertainty. Autistic individuals score relatively high on their intolerance of uncertainty compared to typically developing individuals [6, 15, 67]. Studies often report that the initial interactions with a robot can be stressful for some of these individuals and lead to an aversive reaction to the robot [e.g., References 88, 103]. However, even autistic individuals find a certain degree of unpredictability permissible, or even desirable, as a low-dose of novelty and surprise can be motivating for autistic children, rather than cause distress [2]. We also expected that intolerance of uncertainty may be related to the warmth and competence associated with the robot, but found no evidence for this relation. Unlike discomfort, which is a dimension unique to people's social perception of

robots [13], warmth and competence are argued to be universal dimensions of human social cognition and are used to judge whether another person is a threat or a friend [34]. This means that we found no evidence of a bias of people against our robot in terms of whether the robot intends good or ill.

7.1 Implications

Our findings imply that we have to be careful when designing responsive behaviours for a robot, as the attributed unpredictability may be negatively affected when the cause of the responsive action is not clear. This can hold even when the event is visible, as the person interacting with the robot may not perceive it, or perceive it differently. In turn, the attributed unpredictability can lead to a negative view of the robot's competence, as is indicated by our results. In human-human interaction, people actively improve the behavioural predictability of their actions by using additional signalling actions expressed through a gesture, deictics, facial expressions, body language and posture [17, 51]. Similarly, implementing signalling actions [94, 95] or adding transparency [14] in robot behaviour may be beneficial. These additional actions take time and effort from the robot as well as cognitive effort from the person interacting with it, and therefore should not be implemented *en masse*. But selectively adding them for responsive robot actions may help in communicating the robot's intent and future responsive actions and thereby maintaining attributed predictability. Additionally, endowing robots with the ability to take the perspective of the person interacting with it into account [e.g., References 68, 98] may help identify when a robot's responsive action may be unpredictable and requires additional actions to make it predictable. For example, when a robot is responding to certain events that it has detected, it may be advisable that the robot first communicates its perception to the user, prior to taking the responsive action, or makes the reason for the response transparent in its responsive action. However, more research is needed to explore whether these designs can mitigate or prevent the negative effect of not seeing the cause of a robot's responsive actions on attributed predictability. We investigated one scenario where a robot can behave unpredictably and threaten the attributed predictability and perceived competence of the robot. Identifying other scenarios that can threaten the attributed predictability requires more research, as well as what design solutions can resolve those situations. Note that findings from Driggs-Campbell and Bajcsy [28] indicate that reduced behavioural predictability does not always lead to reduced perceived competence, thus the results from such studies are also likely to be nuanced.

The above applies to the design of robot actions that may be unpredictable, and how they can then negatively influence the robot's attributed predictability. For unpredictable robot motions, the design considerations are different. While we did not experimentally assess the predictability of robot motion, the framework for robot predictability we presented in Section 2, highlights that in initial interactions the person's expectations play a large role in generating predictions, as the person does not yet have experience interacting with the specific robot (model). Studies from Driggs-Campbell and Bajcsy [28] and Dragan and Srinivasa [26] indicate that conforming to such expectations is important for the attributed predictability. An autonomous car that drove with human-inspired motions was found to increase attributed predictability, compared to a car that drove with more mechanical/robot-like motions [28]. As people are already familiar with cars driven by humans, they may also expect similar motions from robot-driven cars. Similarly, a humanoid robot that moves more naturally (i.e., more like a human would) was also found to increase attributed predictability [26]. We expect there to be a correlation between behavioural and attributed predictability, and thus improving the behavioural predictability may also improve the attributed predictability. In human-human interaction, people also actively improve the predictability of their motions by improving the legibility [73] or by being less variable in their motions [100].

Likewise, making robot motions better at expressing intent can also improve their behavioural predictability [e.g., References 27, 40, 93].

7.2 Limitations

The participants in this study had to watch a video showing an interaction of another person interacting with a robot. While this allowed us to carefully control what perceptual information was available to participants to make sense of the robot's behaviour, we need to be careful with generalising the reported findings to individuals who are actually engaging with a robot. In particular, the explored conditions may have different effects on how the robot is perceived when the participant is actually engaged with the robot. Furthermore, while we asked participants to rate the *robot's* unpredictability, and their social perception of the *robot*, we cannot be entirely certain that the participants based their attributed unpredictability solely on the robot's behaviour. The scene itself may also have induced a feeling of unpredictability that was then attributed to the robot. And last, the new scale for measuring robot unpredictability would benefit from being further validated. Our goal was to develop a multi-item scale for measuring the attributed predictability that we could use in the main study, as such scales are less prone to random measurement errors and biases in the meaning and interpretation than is a single item. However, we did not assess whether our new scale is more sensitive and specific than measuring attributed predictability through a single-item, and thus cannot claim that our scale has better validity. Nevertheless, given our assessment of the scale, we felt confident that the scale has sufficient validity and reliability to be used for our main study.

8 CONCLUSION

The predictability of a robot is often mentioned in HRI as an important quality of the robot [e.g., References 1, 43, 45, 52, 58, 69, 85], yet the current conceptual understanding of robot predictability is inadequate as the concept is multi-faceted, rather than a singular concept. This limits us in effectively taking robot predictability into account in the design of robot behaviour. The aim of the study reported in this article was to improve our conceptual understanding of predictability and robots, how robot predictability relates to people's social perception of a robot, and to what extent this is influenced by a person's intolerance of uncertainty. To that end, we presented a framework of what we believe robot predictability is, and we carried out an experimental study. In this study, participants would watch a human interacting with a robot, but we limited the information available to the participants in relation to the robot's responsive actions, to make it more difficult to learn to predict the robot's behaviour. We conclude that not seeing what a robot responds to, negatively influences the attributed predictability and the competence attributed to the robot. The negative effect of not seeing the cause of the robot's responsive actions on competence was largely mediated by the attributed predictability.

We consider our results as preliminary evidence that the predictability attributed to a robot is an important factor to take into account when designing robot behaviour. We have presented our view on the concept of predictability and how it relates to HRI, which may be helpful in understanding and improving robot predictability, and laid out design several potential options on how this could be achieved. Key is to take the attributed predictability of a robot as an important outcome measure that can be used to evaluate different designs (e.g., as in Walker et al. [102]), and can lead to human-robot interactions that are easier to understand. The measurement scale presented in this article showed good qualities in our study and might serve this purpose of measuring the attributed unpredictability of a robot. While we cannot draw conclusions regarding the behavioural predictability of a robot—due to a ceiling effect—we do believe that both behavioural and attributed predictability are important for effective HRI, given the central role of predictability in

human perception. Our novel conceptualisation of predictability in the context of HRI could help in directing future research on predictability so that we may better understand this concept. Since the term “predictability” is often used in HRI literature and it is such a broad concept, we argue that we would do well to clearly communicate what aspect of predictability was studied. Hopefully the framework on robot predictability that we presented, and the distinction that we make between behavioural and attributed predictability, could help us to communicate more clearly about the concept of robot predictability.

In future work, we should look at further investigating the relationship between behavioural and attributed predictability. Furthermore, as predicting the robot's behaviour can be learned, investigating this relationship as it evolves over time will be interesting. While people can learn to predict unnatural robot behaviours (at least to a certain degree [26]), or initially seemingly random robot behaviour, this does not mean that the robot is also considered to be more predictable. As this study was conducted in the broader context of a project that looks to develop robot behaviour for a social robot to support autistic children—who are thought to have difficulties dealing with unpredictability [54, 72, 99] and may benefit from highly predictable interactions with a robot [e.g., References 19, 20, 29, 48, 82, 97]—we are interested in the effect of robot predictability on the engagement of autistic children.

ACKNOWLEDGMENTS

We thank Bridgette Connell and Lynn Packwood for their contribution to this work.

REFERENCES

- [1] Rachid Alami, A. Albu-Schaeffer, A. Bicchi, R. Bischoff, R. Chatila, A. De Luca, A. De Santis, G. Giralt, J. Guiochet, G. Hirzinger, F. Ingrand, V. Lippiello, R. Mattone, D. Powell, S. Sen, B. Siciliano, G. Tonietti, and L. Villani. 2006. Safe and dependable physical human-robot interaction in anthropic domains: State of the art and challenges. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, Beijing, China, 1–16. <https://doi.org/10.1109/IROS.2006.6936985>
- [2] Alyssa M. Alcorn, Helen Pain, and Judith Good. 2014. Motivating children's initiations with novelty and surprise. In *Proceedings of the Conference on Interaction Design and Children (IDC'14)*. ACM Press, Aarhus, Denmark, 225–228. <https://doi.org/10.1145/2593968.2610458>
- [3] Dana H. Ballard, Geoffrey E. Hinton, and Terrence J. Sejnowski. 1983. Parallel visual computation. *Nature* 306, 5938 (Nov. 1983), 21–26. <https://doi.org/10.1038/306021a0>
- [4] Patrizia Basili, Markus Huber, Omiros Kourakos, Tamara Lorenz, Thomas Brandt, Sandra Hirche, and Stefan Glasauer. 2012. Inferring the goal of an approaching agent: A human-robot study. In *Proceedings of the 21st IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'12)*. IEEE, Paris, France, 527–532. <https://doi.org/10.1109/ROMAN.2012.6343805>
- [5] Howard Berenbaum, Keith Bredemeier, and Renee J. Thompson. 2008. Intolerance of uncertainty: Exploring its dimensionality and associations with need for cognitive closure, psychopathology, and personality. *J. Anxiety Disord.* 22, 1 (Jan. 2008), 117–125. <https://doi.org/10.1016/j.janxdis.2007.01.004>
- [6] Christina Boulter, Mark H. Freeston, Mikle South, and Jacqui Rodgers. 2014. Intolerance of uncertainty as a framework for understanding anxiety in children and adolescents with autism spectrum disorders. *J. Autism. Develop. Disorders* 44, 6 (2014), 1391–1402. <https://doi.org/10.1007/s10803-013-2001-x>
- [7] Elliot C. Brown and Martin Brüne. 2012. The role of prediction in social neuroscience. *Front. Hum. Neurosci.* 6 (May 2012), 1–19. <https://doi.org/10.3389/fnhum.2012.00147>
- [8] Harriet Brown, Karl Friston, and Sven Bestmann. 2011. Active inference, attention, and motor preparation. *Front. Psychol.* 2 (Sep. 2011), 1–10. <https://doi.org/10.3389/fpsyg.2011.00218>
- [9] Baptiste Busch, Jonathan Grizou, Manuel Lopes, and Freek Stulp. 2017. Learning legible motion from Human–Robot interactions. *Int. J. Soc. Robot.* 9, 5 (2017), 765–779. <https://doi.org/10.1007/s12369-017-0400-4>
- [10] R. Nicholas Carleton. 2016. Into the unknown: A review and synthesis of contemporary models involving uncertainty. *J. Anxiety Disord.* 39 (Apr. 2016), 30–43. <https://doi.org/10.1016/j.janxdis.2016.02.007>
- [11] R. Nicholas Carleton, Myriah K. Mulvogue, Michel A. Thibodeau, Randi E. McCabe, Martin M. Antony, and Gordon J. G. Asmundson. 2012. Increasingly certain about uncertainty: Intolerance of uncertainty across anxiety and depression. *J. Anxiety Disord.* 26, 3 (2012), 468–479. <https://doi.org/10.1016/j.janxdis.2012.01.011>

- [12] R. Nicholas Carleton, M. A. Peter J. Norton, and Gordon J. G. Asmundson. 2007. Fearing the unknown: A short version of the intolerance of uncertainty scale. *J. Anxiety Disord.* 21, 1 (2007), 105–117. <https://doi.org/10.1016/j.janxdis.2006.03.014>
- [13] Colleen M. Carpinella, Alisa B. Wyman, Michael A. Perez, and Steven J. Stroessner. 2017. The robotic social attributes scale (RoSAS). In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI'17)*. ACM Press, New York, New York, 254–262. <https://doi.org/10.1145/2909824.3020208>
- [14] Joyce Y. Chai, Lanbo She, Rui Fang, Spencer Ottarson, Cody Little, Changsong Liu, and Kenneth Hanson. 2014. Collaborative effort towards common ground in situated human-robot dialogue. In *Proceedings of the ACM/IEEE International Conference on Human-robot Interaction (HRI'14)*. ACM, Bielefeld, Germany, 33–40. <https://doi.org/10.1145/2559636.2559677>
- [15] Paul D. Chamberlain, Jacqui Rodgers, Michael J. Crowley, Sarah E. White, Mark H. Freeston, and Mickle South. 2013. A potentiated startle study of uncertainty and contextual anxiety in adolescents diagnosed with autism spectrum disorder. *Molec. Autism* 4, 1 (2013), 31. <https://doi.org/10.1186/2040-2392-4-31>
- [16] Andy Clark. 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* 36, 3 (June 2013), 181–204. <https://doi.org/10.1017/S0140525X12000477>
- [17] Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge, UK. <https://doi.org/10.1017/CBO9780511620539>
- [18] P. R. Corlett, J. R. Taylor, X.-J. Wang, P. C. Fletcher, and J. H. Krystal. 2010. Toward a neurobiology of delusions. *Progr. Neurobiol.* 92, 3 (Nov. 2010), 345–369. <https://doi.org/10.1016/j.pneurobio.2010.06.007>
- [19] Kerstin Dautenhahn. 1999. Robots as social actors: Aurora and the case of autism. In *Proceedings of the 3rd Cognitive Technology Conference (CT'99)*, Vol. 359. M.I.N.D. Lab, San Francisco, CA, 359–374. Retrieved from <https://books.google.nl/books?id=QM--tgAACAAJ>.
- [20] Kerstin Dautenhahn and Iain Werry. 2004. Towards interactive robots in autism therapy: Background, motivation and challenges. *Pragmat. Cogn.* 12, 1 (2004), 1–35. <https://doi.org/10.1075/pc.12.1.03dau>
- [21] Peter Dayan, Geoffrey E. Hinton, Radford M. Neal, and Richard S. Zemel. 1995. The Helmholtz machine. *Neural Comput.* 7, 5 (Sep. 1995), 889–904. <https://doi.org/10.1162/neco.1995.7.5.889>
- [22] Maartje M. A. de Graaf and Bertram F. Malle. 2019. People's explanations of robot behavior subtly reveal mental state inferences. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI'19)*. IEEE, Daegu, South Korea, 239–248. <https://doi.org/10.1109/HRI.2019.8673308>
- [23] Anca D. Dragan, Shira Bauman, Jodi Forlizzi, and Siddhartha S. Srinivasa. 2015. Effects of robot motion on human-robot collaboration. In *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI'15)*, Vol. 1. ACM Press, New York, New York, 51–58. <https://doi.org/10.1145/2696454.2696473>
- [24] Anca D. Dragan, Kenton C. T. Lee, and Siddhartha S. Srinivasa. 2013. Legibility and predictability of robot motion. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI'13)*. IEEE, Tokyo, Japan, 301–308. <https://doi.org/10.1109/HRI.2013.6483603>
- [25] Anca D. Dragan and Siddhartha S. Srinivasa. 2013. Generating legible motion. In *Robotics: Science and Systems IX*. Robotics: Science and Systems Foundation, Berlin, Germany, 92–103. <https://doi.org/10.15607/RSS.2013.IX.024>
- [26] Anca D. Dragan and Siddhartha S. Srinivasa. 2014. Familiarization to robot motion. In *Proceedings of the ACM/IEEE International Conference on Human-robot Interaction (HRI'14)*. ACM Press, New York, New York, 366–373. <https://doi.org/10.1145/2559636.2559674>
- [27] Anca D. Dragan and Siddhartha S. Srinivasa. 2014. Integrating human observer inferences into robot motion planning. *Auton. Robots* 37, 4 (Dec. 2014), 351–368. <https://doi.org/10.1007/s10514-014-9408-x>
- [28] Katherine Driggs-Campbell and Ruzena Bajcsy. 2016. Communicating intent on the road through human-inspired control schemes. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'16)*. IEEE, Daejeon, South Korea, 3042–3047. <https://doi.org/10.1109/IROS.2016.7759471>
- [29] Audrey Duquette, François Michaud, and Henri Mercier. 2008. Exploring the use of a mobile robot as an imitation agent with children with low-functioning autism. *Auton. Robots* 24, 2 (Feb. 2008), 147–157. <https://doi.org/10.1007/s10514-007-9056-5>
- [30] Friederike Eyssel and Dieta Kuchenbrandt. 2011. Manipulating anthropomorphic inferences about NAO: The role of situational and dispositional aspects of effectance motivation. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'11)*. IEEE, Atlanta, GA, 467–472. <https://doi.org/10.1109/ROMAN.2011.6005233>
- [31] Friederike Eyssel, Dieta Kuchenbrandt, and Simon Bobinger. 2011. Effects of anticipated human-robot interaction and predictability of robot behavior on perceptions of anthropomorphism. In *Proceedings of the 6th International Conference on Human-Robot Interaction (HRI'11)*. ACM Press, Lausanne, Switzerland, 61–67. <https://doi.org/10.1145/1957656.1957673>

- [32] Harriet Feldman and Karl J. Friston. 2010. Attention, Uncertainty, and Free-Energy. *Front. Hum. Neurosci.* 4 (Dec. 2010), 1–23. <https://doi.org/10.3389/fnhum.2010.00215>
- [33] Heike Felzmann, Eduard Fosch-Villaronga, Christoph Lutz, and Aurelia Tamo-Larrieux. 2019. Robots and Transparency: The Multiple Dimensions of Transparency in the Context of Robot Technologies. *IEEE Robot. Autom. Mag.* 26, 2 (6 2019), 71–78. <https://doi.org/10.1109/MRA.2019.2904644>
- [34] Susan T. Fiske, Amy J. C. Cuddy, and Peter Glick. 2007. Universal dimensions of social cognition: warmth and competence. *Trends Cogn. Sci.* 11, 2 (Feb. 2007), 77–83. <https://doi.org/10.1016/j.tics.2006.11.005>
- [35] Paul C. Fletcher and Chris D. Frith. 2009. Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Rev. Neurosci.* 10, 1 (Jan. 2009), 48–58. <https://doi.org/10.1038/nrn2536>
- [36] Amos Freedy, Ewart DeVisser, Gershon Weltman, and Nicole Coeyman. 2007. Measurement of trust in human-robot collaboration. In *Proceedings of the International Symposium on Collaborative Technologies and Systems*. IEEE, Orlando, FL, 106–114. <https://doi.org/10.1109/CTS.2007.4621745>
- [37] Karl Friston. 2003. Learning and inference in the brain. *Neural Netw.* 16, 9 (Nov. 2003), 1325–1352. <https://doi.org/10.1016/j.neunet.2003.06.005>
- [38] Karl J. Friston. 2005. A theory of cortical responses. *Philos. Trans. Roy. Soc. B: Biol. Sci.* 360, 1456 (Apr. 2005), 815–836. <https://doi.org/10.1098/rstb.2005.1622>
- [39] Karl J. Friston. 2010. The free-energy principle: A unified brain theory? *Nature Rev. Neurosci.* 11, 2 (Feb. 2010), 127–138. <https://doi.org/10.1038/nrn2787>
- [40] Michael J. Gielniak and Andrea L. Thomaz. 2011. Generating anticipation in robot motion. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'11)*. IEEE, Atlanta, GA, 449–454. <https://doi.org/10.1109/ROMAN.2011.6005255>
- [41] Richard Langton Gregory. 1968. Perceptual illusions and brain models. *Proc. Roy. Soc. London. Ser. B. Biol. Sci.* 171, 1024 (Dec. 1968), 279–296. <https://doi.org/10.1098/rspb.1968.0071>
- [42] Dan W. Grupe and Jack B. Nitschke. 2013. Uncertainty and anticipation in anxiety: An integrated neurobiological and psychological perspective. *Nature Rev. Neurosci.* 14, 7 (July 2013), 488–501. <https://doi.org/10.1038/nrn3524>
- [43] Peter A. Hancock, Deborah R. Billings, Kristin E. Schaefer, Jessie Y. C. Chen, Ewart J. de Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Hum. Factors: J. Hum. Factors Ergonom. Soc.* 53, 5 (2011), 517–527. <https://doi.org/10.1177/0018720811417254>
- [44] Andrew F. Hayes. 2017. *Introduction to Mediation, Moderation, and Conditional Process Analysis* (2nd ed.). The Guilford Press, New York, New York. Retrieved from <https://www.guilford.com/books/Introduction-to-Mediation-Moderation-and-Conditional-Process-Analysis/Andrew-Hayes/9781462534654>.
- [45] Jochen Heinzmann and Alexander Zelinsky. 2003. Quantitative safety guarantees for physical human-robot interaction. *Int. J. Robot. Res.* 22, 7–8 (July 2003), 479–504. <https://doi.org/10.1177/02783649030227004>
- [46] H. von Helmholtz. 1860. *Handbuch Der Physiologischen Optik* (vol. 3). J. P. C. Southall (Ed.), Dover, New York, NY.
- [47] Jakob Hohwy. 2013. *The Predictive Mind*. Oxford University Press, Oxford, UK. <https://doi.org/10.1093/acprof:oso/9780199682737.001.0001>
- [48] Bibi E. B. M. Huskens, Rianne Verschuur, Jan C. C. Gillesen, Robert Didden, and Emilia I. Barakova. 2013. Promoting question-asking in school-aged children with autism spectrum disorders: Effectiveness of a robot intervention compared to a human-trainer intervention. *Develop. Neurorehab.* 16, 5 (2013), 345–356. <https://doi.org/10.3109/17518423.2012.739212>
- [49] Quentin J. M. Huys, Nathaniel D. Daw, and Peter Dayan. 2015. Depression: A decision-theoretic analysis. *Annu. Rev. Neurosci.* 38, 1 (July 2015), 1–23. <https://doi.org/10.1146/annurev-neuro-071714-033928>
- [50] James M. Kilner, Karl J. Friston, and Christopher D. Frith. 2007. Predictive coding: An account of the mirror neuron system. *Cogn. Process.* 8, 3 (Sep. 2007), 159–166. <https://doi.org/10.1007/s10339-007-0170-2>
- [51] Gary Klein, Paul J. Feltovich, Jeffrey M. Bradshaw, and David D. Woods. 2005. Common ground and coordination in joint activity. In *Organizational Simulation*. John Wiley & Sons, Inc., Hoboken, NJ, 139–184. <https://doi.org/10.1002/0471739448.ch6>
- [52] Markus Koppenborg, Peter Nickel, Birgit Naber, Andy Lungfiel, and Michael Huelke. 2017. Effects of movement speed and predictability in human-robot collaboration. *Hum. Factors Ergonom. Manufactur. Serv. Ind.* 27, 4 (July 2017), 197–209. <https://doi.org/10.1002/hfm.20703>
- [53] Johan Kwisthout, Harold Bekkering, and Iris van Rooij. 2017. To be precise, the details don't matter: On predictive processing, precision, and level of detail of predictions. *Brain Cogn.* 112 (Mar. 2017), 84–91. <https://doi.org/10.1016/j.bandc.2016.02.008>
- [54] Rebecca P. Lawson, Geraint Rees, and Karl J. Friston. 2014. An aberrant precision account of autism. *Front. Hum. Neurosci.* 8, May (May 2014), 302. <https://doi.org/10.3389/fnhum.2014.00302>
- [55] Min Kyung Lee, Jodi Forlizzi, Sara Kiesler, Maya Cakmak, and Siddhartha S. Srinivasa. 2011. Predictability or adaptivity? In *Proceedings of the 6th International Conference on Human-robot Interaction (HRI'11)*. ACM Press, Lausanne, Switzerland, 179. <https://doi.org/10.1145/1957656.1957720>

- [56] Tai Sing Lee and D. Mumford. 2003. Hierarchical Bayesian inference in the visual cortex. *J. Optical Soc. Amer. A* 20, 7 (July 2003), 1434. <https://doi.org/10.1364/JOSAA.20.001434>
- [57] Séverin Lemaignan, Julia Fink, Francesco Mondada, and Pierre Dillenbourg. 2015. Youâ..re doing it wrong! studying unexpected behaviors in child-robot interaction. In *Proceedings of the International Conference on Social Robotics*, Adriana Tapus, Elisabeth André, Jean-Claude Martin, François Ferland, and Mehdi Ammi (Eds.). Lecture Notes in Computer Science, Vol. 9388. Springer International Publishing, Cham, 390–400. <https://doi.org/10.1007/978-3-319-25554-539>
- [58] Michael Lewis, Katia Sycara, and Phillip Walker. 2018. The role of trust in human-robot interaction. In *Studies in Systems, Decision and Control*, 117. Springer International Publishing, Cham, 135–159. <https://doi.org/10.1007/978-3-319-64816-38>
- [59] Christina Lichtenthäler and Alexandra Kirsch. 2014. Goal-predictability vs. trajectory-predictability. In *Proceedings of the ACM/IEEE International Conference on Human-robot Interaction (HRI'14)*. ACM Press, New York, NY, 228–229. <https://doi.org/10.1145/2559636.2559802>
- [60] Christina Lichtenthäler, Tamara Lorenz, Michael Karg, and Alexandra Kirsch. 2012. Increasing perceived value between human and robots—Measuring legibility in human aware navigation. In *Proceedings of the IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO'12)*. IEEE, Munich, Germany, 89–94. <https://doi.org/10.1109/ARSO.2012.6213405>
- [61] Christina Lichtenthäler, Tamara Lorenz, and Alexandra Kirsch. 2012. Influence of legibility on perceived safety in a virtual human-robot path crossing task. In *Proceedings of the 21st IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'12)*. IEEE, Paris, France, 676–681. <https://doi.org/10.1109/ROMAN.2012.6343829>
- [62] Herbert W. Marsh. 1996. Positive and negative global self-esteem: A substantively meaningful distinction or artifacts? *J. Personal. Soc. Psychol.* 70, 4 (1996), 810–819. <https://doi.org/10.1037/0022-3514.70.4.810>
- [63] Merriam-Webster. 2018. Merriam-Webster Thesaurus, accessed September 10th, 2018. Retrieved from <https://www.merriam-webster.com/>.
- [64] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artific. Intell.* 267 (Feb. 2019), 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
- [65] Omar Mubin and Christoph Bartneck. 2015. Do as I say: Exploring human response to a predictable and unpredictable robot. In *Proceedings of the British HCI Conference (HCI'15)*. ACM Press, Lincoln, UK, 110–116. <https://doi.org/10.1145/2783446.2783582>
- [66] Louise Neil, Giulia Cappagli, Themelis Karaminis, Rob Jenkins, and Elizabeth Pellicano. 2016. Recognizing the same face in different contexts: Testing within-person face recognition in typical development and in autism. *J. Exper. Child Psychol.* 143 (Mar. 2016), 139–153. <https://doi.org/10.1016/j.jecp.2015.09.029>
- [67] Louise Neil, Nora Choque Olsson, and Elizabeth Pellicano. 2016. The relationship between intolerance of uncertainty, sensory sensitivities, and anxiety in autistic and typically developing children. *J. Autism Dev. Disorders* 46, 6 (2016), 1962–1973. <https://doi.org/10.1007/s10803-016-2721-9>
- [68] Stefanos Nikolaidis, Anca D. Dragan, and Siddhartha S. Srinivasa. 2016. Viewpoint-based legibility optimization. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI'16)*. IEEE, Christchurch, New Zealand, 271–278. <https://doi.org/10.1109/HRI.2016.7451762>
- [69] Merel Noorman and Deborah G. Johnson. 2014. Negotiating autonomy and responsibility in military robots. *Ethics Info. Technol.* 16, 1 (Mar. 2014), 51–62. <https://doi.org/10.1007/s10676-013-9335-0>
- [70] Marte Otten, Anil K. Seth, and Yair Pinto. 2017. A social Bayesian brain: How social knowledge can shape visual perception. *Brain Cogn.* 112 (Mar. 2017), 69–77. <https://doi.org/10.1016/j.bandc.2016.05.002>
- [71] Elise Payzan-LeNestour and Peter Bossaerts. 2011. Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS Comput. Biol.* 7, 1 (Jan. 2011), e1001048. <https://doi.org/10.1371/journal.pcbi.1001048>
- [72] Elizabeth Pellicano and David Burr. 2012. When the world becomes “too real”: A Bayesian explanation of autistic perception. *Trends Cogn. Sci.* 16, 10 (2012), 504–510. <https://doi.org/10.1016/j.tics.2012.08.009>
- [73] Giovanni Pezzulo, Francesco Donnarumma, and Haris Dindo. 2013. Human sensorimotor communication: A theory of signaling in online social interactions. *PLoS ONE* 8, 11 (Nov. 2013), e79876. <https://doi.org/10.1371/journal.pone.0079876>
- [74] W. James Potter and Deborah Levine-Donnerstein. 1999. Rethinking validity and reliability in content analysis. *J. Appl. Commun. Res.* 27, 3 (8 1999), 258–284. <https://doi.org/10.1080/00909889909365539>
- [75] David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behav. Brain Sci.* 1, 4 (Dec. 1978), 515. <https://doi.org/10.1017/S0140525X00076512>

- [76] Marco Ragni, Andrey Rudenko, Barbara Kuhnert, and Kai O. Arras. 2016. Errare humanum est: Erroneous robots in human-robot interaction. In *Proceedings of the 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'16)*. IEEE, New York, New York, 501–506. <https://doi.org/10.1109/ROMAN.2016.7745164>
- [77] Rajesh P. N. Rao and Dana H. Ballard. 1999. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neurosci.* 2, 1 (Jan. 1999), 79–87. <https://doi.org/10.1038/4580>
- [78] Thomas L. Rodebaugh, Carol M. Woods, David M. Thissen, Richard G. Heimberg, Dianne L. Chambless, and Ronald M. Rapee. 2004. More information from fewer questions: The factor structure and item properties of the original and brief fear of negative evaluation scale. *Psychol. Assess.* 16, 2 (2004), 169–181. <https://doi.org/10.1037/1040-3590.16.2.169>
- [79] Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2013. To err is human(-like): Effects of robot gesture on perceived anthropomorphism and likability. *Int. J. Soc. Robot.* 5, 3 (2013), 313–323. <https://doi.org/10.1007/s12369-013-0196-9>
- [80] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. 2015. Would you trust a (faulty) robot? In *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI'15)*. ACM Press, New York, New York, 141–148. <https://doi.org/10.1145/2696454.2696497>
- [81] Tamie Salter, Iain Werry, and François Michaud. 2008. Going into the wild in child-robot interaction studies: Issues in social robotic development. *Intell. Serv. Robot.* 1, 2 (Apr. 2008), 93–108. <https://doi.org/10.1007/s11370-007-0009-9>
- [82] Felipe Sartorato, Leon Przybylowski, and Diana K. Sarko. 2017. Improving therapeutic outcomes in autism spectrum disorders: Enhancing social communication and sensory processing through the use of interactive robots. *J. Psych. Res.* 90 (July 2017), 1–11. <https://doi.org/10.1016/j.jpsychires.2017.02.004>
- [83] Sau-lai Lee, Ivy Yee-man Lau, S. Kiesler, and Chi-Yue Chiu. 2005. Human mental models of humanoid robots. In *Proceedings of the IEEE International Conference on Robotics and Automation*, Vol. 2005. IEEE, Barcelona, Spain, 2767–2772. <https://doi.org/10.1109/ROBOT.2005.1570532>
- [84] Bob R. Schadenberg. 2021. *Robots for Autistic Children—Understanding and facilitating predictability for engagement in learning*. Ph.D. Dissertation. University of Twente, Enschede, The Netherlands. <https://doi.org/10.3990/1.9789036551649>
- [85] Alessandra Sciutti, Martina Mara, Vincenzo Tagliasco, and Giulio Sandini. 2018. Humanizing human-robot interaction: On the importance of mutual understanding. *IEEE Technol. Soc. Mag.* 37, 1 (Mar. 2018), 22–29. <https://doi.org/10.1109/MTS.2018.2795095>
- [86] Natalie Sebanz, Harold Bekkering, and Günther Knoblich. 2006. Joint action: Bodies and minds moving together. *Trends Cogn. Sci.* 10, 2 (Feb. 2006), 70–76. <https://doi.org/10.1016/j.tics.2005.12.009>
- [87] Natalie Sebanz and Guenther Knoblich. 2009. Prediction in joint action: What, when, and where. *Topics Cogn. Sci.* 1, 2 (4 2009), 353–367. <https://doi.org/10.1111/j.1756-8765.2009.01024.x>
- [88] Elaine Schaertl Short, Eric C. Deng, David J. Feil-Seifer, and Maja J. Matarić. 2017. Understanding agency in interactions between children with autism and socially assistive robots. *J. Human-Robot Interact.* 6, 3 (Dec. 2017), 21. <https://doi.org/10.5898/JHRI.6.3.Short>
- [89] Elaine Schaertl Short, Justin W. Hart, Michelle Vu, and Brian Scassellati. 2010. No fair!! An interaction with a cheating robot. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI'10)*. IEEE, Osaka, Japan, 219–226. <https://doi.org/10.1109/HRI.2010.5453193>
- [90] Pawan Sinha, Margaret M. Kjelgaard, Tapan K. Gandhi, Kleovoulos Tsourides, Annie L. Cardinaux, Dimitrios Pantazis, Sidney P. Diamond, and Richard M. Held. 2014. Autism as a disorder of prediction. *Proc. Natl. Acad. Sci. U.S.A.* 111, 42 (2014), 15220–15225. <https://doi.org/10.1073/pnas.1416797111>
- [91] Paul E. Spector, Paul T. Van Katwyk, Michael T. Brannick, and Peter Y. Chen. 1997. When two factors don't reflect two constructs: How item characteristics can produce artifactual factors. *J. Manage.* 23, 5 (1997), 659–677. <https://doi.org/10.1177/014920639702300503>
- [92] Sonja Stange and Stefan Kopp. 2020. Effects of a social robot's self-explanations on how humans understand and evaluate its behavior. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. ACM, New York, NY, 619–627. <https://doi.org/10.1145/3319502.3374802>
- [93] Daniel Szafrir, Bilge Mutlu, and Terrence Fong. 2014. Communication of intent in assistive free flyers. In *Proceedings of the ACM/IEEE International Conference on Human-robot Interaction (HRI'14)*. ACM Press, Bielefeld, Germany, 358–365. <https://doi.org/10.1145/2559636.2559672>
- [94] Daniel Szafrir, Bilge Mutlu, and Terry Fong. 2015. Communicating directionality in flying robots. In *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI'15)*. ACM Press, New York, New York, 19–26. <https://doi.org/10.1145/2696454.2696475>
- [95] Leila Takayama, Doug Dooley, and Wendy Ju. 2011. Expressing thought: Improving robot readability with animation principles. In *Proceedings of the 6th International Conference on Human-robot Interaction (HRI'11)*. ACM Press, Lausanne, Switzerland, 69. <https://doi.org/10.1145/1957656.1957674>

- [96] Diana I. Tamir and Mark A. Thornton. 2018. Modeling the predictive social mind. *Trends Cogn. Sci.* 22, 3 (Mar. 2018), 201–212. <https://doi.org/10.1016/j.tics.2017.12.005>
- [97] Serge Thill, Cristina A. Pop, Tony Belpaeme, Tom Ziemke, and Bram Vanderborght. 2012. Robot-assisted therapy for autism spectrum disorders with (partially) autonomous control: Challenges and outlook. *Paladyn, J. Behav. Robot.* 3, 4 (Jan. 2012), 209–217. <https://doi.org/10.2478/s13230-013-0107-7>
- [98] J. G. Trafton, N. L. Cassimatis, M. D. Bugajska, D. P. Brock, F. E. Mintz, and Alan C. Schultz. 2005. Enabling effective human–robot interaction using perspective-taking in robots. *IEEE Trans. Syst., Man Cybernet. Part A: Syst. Hum.* 35, 4 (July 2005), 460–470. <https://doi.org/10.1109/TSMCA.2005.850592>
- [99] Sander Van de Cruys, Kris Evers, Ruth Van der Hallen, Lien Van Eylen, Bart Boets, Lee De-Wit, and Johan Wagemans. 2014. Precise minds in uncertain worlds: Predictive coding in autism. *Psychol. Rev.* 121, 4 (2014), 649–675. <https://doi.org/10.1037/a0037665>
- [100] Cordula Vesper, Robrecht P. R. D. van der Wel, Günther Knoblich, and Natalie Sebanz. 2011. Making oneself predictable: Reduced temporal variability facilitates joint action coordination. *Exper. Brain Res.* 211, 3–4 (June 2011), 517–530. <https://doi.org/10.1007/s00221-011-2706-z>
- [101] Catherine Wacongne, Etienne Labyt, V. van Wassenhove, Tristan Bekinschtein, Lionel Naccache, and Stanislas Dehaene. 2011. Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proc. Natl. Acad. Sci. U.S.A.* 108, 51 (Dec. 2011), 20754–20759. <https://doi.org/10.1073/pnas.1117807108>
- [102] Phillip Walker, Michael Lewis, and Katia Sycara. 2016. The effect of display type on operator prediction of future swarm states. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC'16)*. IEEE, Budapest, Hungary, 002521–002526. <https://doi.org/10.1109/SMC.2016.7844619>
- [103] Zachary E. Warren, Zhi Zheng, Shuvajit Das, Eric M. Young, Amy R. Swanson, Amy F. Weitlauf, and Nilanjan Sarkar. 2015. Brief report: Development of a robotic intervention platform for young children with ASD. *J. Autism Dev. Disorders* 45, 12 (Dec. 2015), 3870–3876. <https://doi.org/10.1007/s10803-014-2334-0>
- [104] Bogdan Wojciszke, Roza Bazinska, and Marcin Jaworski. 1998. On the dominance of moral categories in impression formation. *Personal. Soc. Psychol. Bull.* 24, 12 (Dec. 1998), 1251–1263. <https://doi.org/10.1177/01461672982412001>
- [105] Valentin Wyart, Anna Christina Nobre, and Christopher Summerfield. 2012. Dissociable prior influences of signal probability and relevance on visual contrast sensitivity. *Proc. Natl. Acad. Sci. U.S.A.* 109, 9 (Feb. 2012), 3593–3598. <https://doi.org/10.1073/pnas.1120118109>
- [106] Angela J. Yu and Peter Dayan. 2005. Uncertainty, neuromodulation, and attention. *Neuron* 46, 4 (May 2005), 681–692. <https://doi.org/10.1016/j.neuron.2005.04.026>
- [107] Xuan Zhao, Corey Cusimano, and Bertram F. Malle. 2016. Do people spontaneously take a robot’s visual perspective? In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI'16)*. IEEE, New York, New York, 335–342. <https://doi.org/10.1109/HRI.2016.7451770>

Received December 2019; revised April 2021; accepted April 2021