The Perception of Agency

J. GREGORY TRAFTON, Naval Research Laboratory, USA
J. MALCOLM MCCURRY, Arcfield, USA
KEVIN ZISH, Global Systems Technologies, USA
CHELSEA R. FRAZIER, Unites States Military Academy, USA

The perception of agency in human robot interaction has become increasingly important as robots become more capable and more social. There are, however, no accepted or consistent methods of measuring perceived agency; researchers currently use a wide range of techniques and surveys. We provide a definition of perceived agency and from that definition we create and psychometrically validate a scale to measure perceived agency. We then perform a scale evaluation by comparing the PA scale constructed in experiment 1 to two other existing scales. We find that our PA and PA-R (Perceived Agency - Rasch) scales provide a better fit to empirical data than existing measures. We also perform scale validation by showing that our scale shows the hypothesized relationship between perceived agency and morality.

Additional Key Words and Phrases: perceived agency, agency, human robot interaction

ACM Reference Format:
J. Gregory Trafton, J. Malcolm McCurry, Kevin Zish, and Chelsea R. Frazier. 2023. The Perception of Agency. 1, 1 (December 2023), 24 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Does a person who observes or interacts with a robot think it is making its own decisions? Or has it been programmed for that exact situation? These questions are of perceived agency and have implications for design [46], law [4], interaction studies [77], philosophy [10], morality [5] and social psychology [32]. While agency has been well studied, there are also many different overlapping concepts – animacy, mind-perception, anthropomorphism, intentionality, and others. These concepts are similar but distinct from perceived agency. Animacy in robotics focuses on making the robot lifelike, frequently focusing on how the robot moves [6]. Anthropomorphism concerns attributing human characteristics or behavior to a robot or other non-human entity [42]. Mind perception is concerned with how people conceptualize others' minds – the number of dimensions and what those dimensions are [31, 55, 90]. Intentionality is how deliberately and goal-directed a robot acts and is frequently associated with perceived agency [75]. In this report, we focus on perceived agency.

In their seminal work, Gray, Gray, and Wegner (2007) explored how people think about other people's minds [31]. Specifically, they were interested in the number of dimensions that people thought others' minds consisted of. They found, contrary to current beliefs, that people conceptualized others' minds along two dimensions: experience (the extent to which an entity is capable of being hungry, feeling rage, desire, pleasure, pain, etc.) and agency (the extent

Authors' addresses: J. Gregory Trafton, greg.trafton@nrl.navy.mil, Naval Research Laboratory, 4555 Overlook AV SW, Washington, DC, USA, 20375; J. Malcolm McCurry, malcolmmccurry@gmail.com, Arcfield, USA; Kevin Zish, zish.kevin@gmail.com, Global Systems Technologies, USA; Chelsea R. Frazier, chelsea.r.frazier.mil@army.mil, Unites States Military Academy, USA.

2023. Manuscript submitted to ACM

to which an entity is capable of recognizing emotions, having self-control, planning, communication, morality, and thought).

Gray et al. (2007) used a principal component analysis (PCA) to analyze their data. PCA and factor analysis are statistical approaches that reduce the dimensionality of large datasets. For example, Gray et al. found that when individuals answered questions like "Which one do you think is more capable of feeling hungry, a robot or a 5-year-old girl?" on a 5 point Likert scale, some capabilities were answered similarly (e.g., feeling hungry and feeling pain had comparable scores across a range of entities). Some capabilities were associated with each other more frequently than another set. Each set of questions that were strongly correlated with each other but were correlated less with other questions could be considered a dimension or factor. These dimensions are considered latent – not directly observed but inferred from the combination of associated questions.

A decade later, Weisman, Dweck, and Markman (2017) changed the original methodology and suggested that instead of 2 dimensions of mind perception, there were 3 [90]. Weisman et al. argued that the three dimensions of mind perception are body (e.g., getting hungry, experiencing pain, feeling tired), heart (e.g., feeling love, having a personality), and mind (e.g., remembering things, detecting sounds). Both the experience and agency concepts from [31] were scattered across all [90]'s body, heart, and mind dimensions.

Malle (2019) used a similar methodology to [90], but used different initial items [55]. Interestingly, Malle also found three dimensions, but they were slightly different than [90] and in some cases showed five dimensions. Malle found Affect (positive and negative emotions and feelings), moral (e.g., telling right from wrong) / social cognition (planning and theory of mind), and reality interaction (verbal communication and moving through the environment). Again, both the experience and agency concepts loaded on different dimensions. All three of these studies used a strong bottom-up approach and to search for items that were associated with mind perception.

Because both Weisman et al. (2017) and Malle (2019) did not find evidence that agency was one of the core dimensions of mind perception, other researchers have been understandably uncertain about the status of perceived agency and how to measure it. We believe that while agency is not a core dimension of mind perception, it can be measured as a component of how people perceive other entities, much like the Robotic Social Attribute Scale (RoSAS) measures warmth, competence, and discomfort of robots [11] or how the multidimensional measure of trust (MDMT) measures different dimensions of trust [87].

Previous experimental work in perceived agency has focused on determining whether non-organic entities (i.e., robots, AI characters) can be perceived as having agency and what cues lead people to judge whether an entity has agency. Multiple researchers have shown that people do, in fact, ascribe agency to non-organic entities. For example, Heider and Simmel (1944) constructed an animation of geometric shapes and noticed that people frequently ascribed the shapes with goals, emotions, and perceived agency [36, 75]. This work launched an entire subfield investigating how adults and children perceive intentional motion and the relationship to goal-directed cognition and perceived agency [26, 75, 76].

Researchers originally hypothesized that when an entity looks like or acts "like a person," the entity is more likely to be perceived as having agency [42, 57]. Later researchers, however, have attempted to find better and more specific cues over the general "like a person" hypothesis. For example, Short et al. (2010) used a very clever experimental paradigm with a robot playing a game of rock paper scissors [77] to examine perceived agency. In one condition, the robot played in a standard way throughout multiple rounds with a participant. In another condition, the robot seemed to make a mistake when calling out who won or lost. In a final condition, the robot actively cheated by changing their throw after both the robot and the participant had completed the round. The cheating robot was perceived as having more agency than the other two robots.

Another group of researchers have shown that robots with social norms may be a cue that lead people to believe that the robot has agency. For example, Korman et al. (2019) found that robots that follow social norms are perceived as having more agency than robots that disregard social norms or that seem to make a mistake [45]. Yasuda et al. (2020) refined this hypothesis and found that a robot that cheated was perceived as more agentic than other types of social norm violations (cursing or insulting), suggesting that cheating itself may be one of the features that encourages people to think of robots as having agency. [94].

1.1 Measuring Perceived Agency

It is clear from this brief review that a great deal of research has already been done on perceived agency, and a large number of claims have been made about perceived agency. However, this review also masks a serious problem: we do not have a reliable, robust, theoretically meaningful method of measuring perceived agency. This problem can be shown by examining how a number of influential papers from the last few years have measured perceived agency. Some researchers have measured perceived agency through qualitative coding of written comments [77, 94]. Another group of researchers have used overlapping concepts of animacy or anthropomorphism to make claims about perceived agency [6, 91]. A different group of researchers have used idiosyncratic measures of perceived agency where they created measures of perceived agency for their specific study or used incomplete scales from other sources, e.g., [33, 48, 68]; unfortunately, all the idiosyncratic measures were different from each other and each paper made strong claims about perceived agency. Finally, some of these measures show inconsistent results across experiments [54, 77, 94].

This lack of a good measurement tool inhibits our theoretical understanding of what perceived agency is, but also how it impacts other constructs (or vice versa). Because the measurement of perceived agency is so different across studies, the conclusions and opportunities for replication are limited. All these reasons suggest a reliable method of measuring PA is needed to increase theory and practice of HRI. Our goal in this report will be to construct a method for measuring perceived agency in entities of all types.

2 PERCEIVED AGENCY: DEFINITION

The first step in most survey development research is to create a strong conceptual definition [8, 59]; this definition can then be used to construct or select items that are consistent with the definition. We take Dennet (1978) as inspiration [19] and suggest that:

People perceive agency in another entity when the entity's actions may be assumed by an outside observer to be driven primarily by its internal thoughts and feelings and less by the external environment.

The importance of another's thoughts and feelings in the perception of agency has been highlighted before, both in our introduction and by others [31, 40, 75]. We felt that it was important to include a locus of control component in our definition as well. One of the traditional strengths of robots and artificial intelligence agents is that they excel at performing repetitive tasks, but usually only in a specific environment. Locus of control is an individual's perception about the causes of their actions: whether self-generated or caused by external forces [64, 72, 81]. Thus, our definition allows the possibility that the external world could be a possible cause of an entity's actions.

3 GENERATING MEASUREMENT INSTRUMENTS

The most common method of generating and validating a scale is to use factor analysis [25, 69]. The general approach to construct a validated instrument from factor analysis is described in detail by others [8] and the factor analytic approach has had success in HRI [11, 63, 87]. The factor analytic approach to survey construction typically consists of generating a large number of possible items that relate to the dimension of interest. Participants then use those items to rate an entity, an interaction, or themselves. Factor analysis provides loadings that describe how related each item is to different dimensions (factors). Items that highly load on a specific factor can be considered consistent with that factor. Different factors are usually considered different aspects of the primary area of interest.

In the factor analytic approach, items are selected to maximize reliability which leads to items that are similar in terms of endorsability [22, 78]. This is an excellent approach when the researcher is attempting to understand the many dimensions and nuances of the construct (e.g., the mind perception work described earlier). Indeed, Smith (2002) suggests using factor analysis when the data have multiple uncorrelated factors [80].

The factor analytic approach has at least two disadvantages when attempting to create a measurement scale. First, because factor analysis identifies how close an item is to the underlying latent variable, it can be more difficult to select items that cover a wide range of the latent variable. This can make it more difficult to differentiate levels or amounts of the specific dimension the scale is measuring.

Second, the ideal of scale development is to measure a single dimension, or latent factor [17, 60, 71]. A single dimension is desirable because it makes interpretation and understandability easier and more straightforward. When a construct does have multiple constructs, difficulties in interpretation can arise because the analyst needs to show how the multiple factors create a general factor [71]: this occurs more commonly for understanding the factor's structure and less when the focus is scale creation. Factor analysis can measure and show unidimensional constructs, but the number of dimensions in factor analysis is still a hotly debated topic [69].

A final feature about factor analysis is that it measures the latent construct of a person and does not account for an external target. Usually this is not a concern – an individual's attitudes or opinions can be well measured. However, when the target is an external event or entity, factor analysis has no way to account for differences in those external targets.

Because our intention is to construct a unidimensional scale of perceived agency about external entities (e.g., robots or AIs), we will be using a Rasch analysis.

The Rasch model is a mathematical formulation that describes the relationship between raters, items, and entities and is part of the item response theory framework. All three components are measured on the same latent scale which is a logit as the unit of measurement. Rasch analysis models the fact that raters, entities, and items can all vary along the latent variable: in our case, some raters will have a (pre)disposition to believe an entity has more or less perceived agency; an item may be easier or more difficult to agree with; and an entity may have more or less perceived agency. Because Rasch puts all three components on the same measurement scale, it is straightforward to determine the location of each rater, item, or entity.

Rasch models have measurement invariance [9, 21, 23]: when a set of observations fits a Rasch model, entity measures are invariant across different sets of items or raters, and items and raters are invariant across different entities. Measurement invariance suggests that test scores are sufficient statistics for estimating rater measures. Measurement invariance is tested by fit statistics [79]; unidimensionality and reliability can be measured as well.

3.1 Rasch Analysis

Our approach will be to have raters (participants) answer items (survey questions) on different entities that will be judged. Each of these "facets" is a separate source of information and bias and each can be measured along the same dimension. Rasch analsyis can construct measurements for each element in each facet.

Items in a Rasch analysis perform best if there is a range where some items are easier to agree with and some are more difficult to agree with. Each item will be expected to measure some aspect of the latent trait (perceived agency) that we are interested in.

Entities in our case will consist of a variety of videos that show a robot, AI character, or person performing some task. Like items, entities will be expected to have a range of perceived agency.

Raters are people who watch a video and answer items about the entity. A rater who may be more likely to agree that many entities have some amount of perceived agency would be considered to have more of the latent value. Similarly, a rater who felt that very few entities could have perceived agency would be considered to have less of the perceived agency latent value. For example, a person who felt that very few entities have much perceived agency would be scored as having relatively little perceived agency as a latent value. These latent values for each rater can be considered a (pre)disposition for whether an entity may have perceived agency.

We can operationalize these intuitions using a Rasch rating scale, which can be defined as:

$$ln(\frac{P_{eirc}}{P_{eir(c-1)}}) = \theta_e - \beta_i - \alpha_r - \tau_c \tag{1}$$

where

- c is the category of the rating scale or the Likert value; in our case it will be 1-5,
- P_{eirc} is the probability of entity e receiving a rating of category c of item i from rater r,
- $P_{eir(c-1)}$ is the probability of entity e receiving a rating of category $c-1$ of item i from rater r,
- θ_e is the amount of perceived agency of entity e,
- β_i is how difficult the item i is to agree with,
- α_r is the severity or (pre-)disposition of rater r, and
- τ_c is the difficulty of receiving a rating of category c relative to a rating of category $c-1$

The category value, τ_c is the location where adjacent categories, c and $c-1$ are equally probable to be observed, also known as Rasch-Andrich thresholds [51].

The Rasch model is an additive linear model based on a logistic transformation of ratings to a logit scale. Critically, each facet (entities e, raters r, items i) are all on the same logit scale and all can influence the final rating. Conceptually, this means that the logit scale represents the latent value or dimension – the amount of perceived agency.

The Rasch model makes some basic assumptions about measurement [9, 16, 74, 93]. For example, if a rater with a high (pre-)disposition of perceived agency (α) gives an entity with a high perceived agency (θ) an especially low score on an item (β), that item may have have a measurement problem or mis-fit. Rasch models allow us to find and inspect these mis-fits; entities, items, or raters who have consistently large mis-fits suggest a concern: the video may be misleading; an item may be confusing; a rater may be answering randomly. Rasch models have several strengths, including generalizability across entities and raters (e.g., different robots or AIs can be measured accurately by different raters), performs measurements in an interval scale (not an ordinal), allowing parametric statistical analysis, can identify items or entities that do not behave as expected, and produces an ordered set of items and entities.

4 EXPERIMENT 1: SCALE CONSTRUCTION

The goal of experiment 1 was to generate a set of items that could accurately measure perceived agency across a wide range of entities. The focus here will be on robots, but we will also include AI characters and humans.

4.1 Method

All studies, including this one were approved by the NRL IRB. All participants consented to participate.

4.1.1 Participants. The suggested number of participants for a Rasch analysis to achieve a 95%+ confidence of measures within .5 logits is 150 [49]. 195 participants were recruited through Cloud Research and paid $12 for participation in the study. 9 participants were removed because they missed an attention check ("has a face") leaving 186 participants. The average age of participants was 35 (SD=12) years old. 108 participants were women, 77 participants were men, and 1 participant was unreported. The study took 29 minutes on average.

Table 1. Description of videos used.

Label	Group	Entity actions	Morphology	Source
Industrial	1	Stacking and moving boxes	industrial arm	[2]
Feeder	1	Picking up food with a fork and feeding a human	Arm	[37, 38]
Soccer	1	Shooting soccer goals	Humanoid	[18]
Bargaining1	1	Human bargaining with AI agent	Humanoid character	[30]
Cheating RPS	1	Human playing a game of rock paper scissors; robot cheats	Humanoid	[94]
Robot Secrets Revealed	1	Humans test robots who then rebel	Humanoid	[34]
Teacher	1	Algebra math teacher	Human	[92]
Palletizer	2	Stacking and moving pallets	industrial arm	[44]
Dishes	2	Moving coffee cups into strainer	Arm	[67]
Line	2	Reading signs and cutting in line	Rolling Humanoid	[45]
Firefighting	2	Receiving instructions from a human and putting out a fire	Humanoid	[58]
Bargaining2	2	Human bargaining with AI agent	Humanoid character	[30]
Service	2	Helping a human find a sports jersey in a store	Humanoid	[43]
Musician	2	Musician playing 4 parts	Human	[86]

4.1.2 Materials (Videos). 14 Videos were selected and collected from a wide range of sources, including YouTube, academic conference proceedings, and personal communication with leaders of the field in robotics. The majority of our stimuli were robots (10), but also included AI agents (2) and humans (2). The entities portrayed a range of engagement with people (from none to speaking and interacting) and the non-human entities had different morphologies, differed in their sensing, and had different perceptual, navigation, mobility, cognitive, and social capabilities. The videos were divided into two groups of seven that, according to pilot testing, had a comparable range of perceived agency.

Table 1 provides a label, the group the entity was placed in, a brief description, the morphology of the robot, and a citation of the source of the robot. The citation of each video is either a YouTube location or a paper or website describing the video.

Our goal was to keep the videos between 30 seconds and 3 minutes. In some cases the video was trimmed or cut. In all cases, we attempted to show the core aspects of the entity and their activity while making sure that participants would not become bored while watching the video.

4.1.3 Materials (Survey Items). Data collection A and B in the Appendix show initial attempts at item development for perceived agency. Based on those data collection efforts as well as difficulties that participants in those studies mentioned, we created a set of items that captured core aspects of thoughts, feelings, and the impact of the external environment on an entity.

In addition to items that covered thoughts, emotions, and environmental impacts on behavior, we also included two integrative items. The actor scenario was "Imagine the robot/character/person was asked to be an actor in a local theater production. How well do you think they would do?" The dinner scenario was "Imagine the robot/character/person was asked to host a dinner party for your friends next weekend. This includes coming up with a menu, cooking, and hosting. How well do you think they will do?" These integrative items were included to examine whether only a combination of thoughts, features, and environmental impacts on behavior would be able to predict perceived agency.

As mentioned previously, Rasch analysis benefits from having items that range in their difficulty to agree. In this set, there were some items that were on average easy to agree with (i.e., "acts with purpose") to items that were on average more difficult to agree with (i.e., "can show emotions to other people").

A complete list of the survey items used are shown in Table 2. All items were on a five point Likert scale.

Table 2. Survey items used. These questions were prefaced at the top of the column with "The <entity>..."

Order	Focus	Survey Item
1	Thoughts	acts with purpose
2	Thoughts	has goals
3	Thoughts	can create new goals
4	Thoughts	can communicate with people
5	Thoughts	treats others as if they had a mind
6	Feelings	wanted to perform these actions
7	Feelings	can show emotions to other people
8	Feelings	can change their behavior based on how people treat them
9	Environment	can adapt to different situations
10	Environment	would do well in other environments
11	attention check	has a face
12	Environment	can perform many different types of tasks
13	Integrated	actor scenario
14	Integrated	dinner scenario

4.1.4 Procedure. Participants were randomly placed in either group 1 or group 2. After answering a series of demographic questions, participants were given a brief description of the task and told they would answer a series of questions about 7 different videos. For each of the 7 videos, each participant was randomly shown one of the videos from the group they were assigned. At the end of the video, they were taken to a single page with the same video that they could watch again if desired. They were first asked to describe the video in at least one sentence. Next they were asked to answer the survey questions in Table 2 about the entity in the video. A thumbnail of the entity was provided as well to reduce confusion. The words “The robot/character/human” was at the top of the column for the survey questions.

After participants completed all the items, they could advance to the next video, and the entire process repeated for each of the 7 videos. All the Likert questions had to be answered in order to progress to the next video. After the 4th video, the participant was offered a break before continuing. Additionally, to provide pacing for the participant, the number of videos that had been seen and the number remaining were provided (e.g., 4 out of 7).

Finally, at the end of the session, participants were invited to provide experimental feedback.

4.2 Results

We performed a Rasch analysis using Facets version 3.83.6 [52]. Because there were two groups and we wanted to create an integrated scale for both groups and all items, we linked the two groups by assuming the two groups had equal amounts of the latent value. All raters came from the same population and data was collected concurrently for both groups (alternating). Entities are typically ordered from highest to lowest, but other facets are conventionally reversed; here, items and raters have their sign reversed so that items that are more difficult to agree with and raters that are the most lenient have the highest latent value.

The Rasch model will be evaluated by (1) examining the extent of item unidimensionality; (2) examining reliability and separation; and (3) examining fit statistics for entities, raters, and items.

4.2.1 Unidimensional. Using a Rasch analysis for scale construction works best when one latent variable is sufficient to explain most of the variation in the responses. One common way to examine whether the items are measuring a single latent dimension is to perform a principle component analysis (PCA) of the standardized residuals [14]; if the standardized residuals are 2 or more, there is evidence another dimension exists in the data [66]. A PCA of the standardized residuals showed that the eigenvalues for the first contrast was 1.0, suggesting that the residual was noise, not another latent factor. This result suggests that the resulting logit scale was unidimensional.

4.2.2 Reliability. Rasch analysis provides two different measures for reliability. The first, separation, indicates how many different levels can be distinguished. A small separation value suggests that different levels can not be distinguished while a larger value is more desired for measurement. The second reliability measure, separation reliability, is equivalent to cronbach alpha reliability and is a measure of internal consistency. Separation reliability ranges from 0 to 1; over .8 is considered acceptable for scale creation.

The separation reliability value for entities was > 0.95 and the separation index was 31.2. These values indicate that over 30 levels of entities can be distinguished with this scale and that there was very high reliability.

The separation reliability value for raters was .95 and the separation index was 4.3. These values indicate that approximately four levels of raters can be distinguished with this scale and that there was very high reliability. Some raters were much more predisposed to attributing agency to an entity than others.

The separation reliability value for items was > 0.95 and the separation index was 23.9. These values indicate that the ordering of the items is reliable. In aggregate, reliability and separation is quite high for this experiment.

4.2.3 Fit statistics. While traditional factor analysis has a set of methods to determine how well items relate to latent constructs (i.e., loadings, item scale correlations, etc.), Rasch uses different methods. Rasch identifies departures in the data for persons, items and even data points from the ideal of unidimensionality. These are reported with fit statistics that guide the improvement of the instrument and point out possible flaws in the data.

There are two common fit statistics used for Rasch analysis: infit and outfit. Outfit is an unweighted fit statistic, a measure of how well the data fit the model and is the most common method for evaluating Rasch fit, and what we will use here. The outfit statistic is sensitive to large departures from model expectations: if an otherwise high-perceived-agency rater gives a high perceived agency entity a low score, this would show a high outfit and highlight a potential concern. Low outfits signal that there is very little additional information provided. A low outfit is considered $< .5$ while a high outfit is considered > 1.5 [53]. Rasch analysis provides outfits for each facet; in our case, items, entities, and, raters.

Recall that all latent values are on a logit scale with a mean of 0 and a standard deviation of 1. Logit scores have an infinite range but typically range from ± 5. Table 3 shows the modeled latent value for items, β (see equation 1). The higher the value of β, the more difficult it is to agree that an entity has perceived agency. Thus, it is relatively easy to agree that most tested entities act with purpose ($\beta = -1.51$), but it is much more difficult to agree that the tested entities would do well as an actor ($\beta = .90$) or can show emotions to other people ($\beta = .64$). Table 3 also shows the outfit for items. All items are within the acceptable outfit ranges from 0.5 to 1.5 except for "acts with purpose" which has an outfit of exactly 1.5.

Item	β	Outfit	β	Outfit
	Orig	Orig	Revised	Revised
dinner scenario	1.07	1.00	1.16	1.05
actor scenario	.90	.87	.97	.90
can show emotions to other people	.64	.77	.69	.77
can change their behavior based on how people treat them	.36	.87	.37	.86
treats others as if they had a mind	.20	.85	.19	.80
can create new goals	.07	.85	.08	.84
would do well in other environments	.05	1.16	.03	1.13
can perform many different types of tasks	.04	1.20	.04	1.19
can communicate with people	-.11	.92	-.13	.87
can adapt to different situations	-.11	1.05	-.12	1.06
wanted to perform these actions	-.47	1.25	-.49	1.30
has goals	-1.13	1.38	-1.19	1.37
acts with purpose	-1.51	**1.5**	-1.59	1.31

Table 3. Item βs and outfits. All β SEs are ≤ 0.04. Original are results with Musician; Revised are results after Musician was removed from the analysis.

Table 4 shows the modeled latent value for entities, θ (see equation 1). The higher the value of θ, the more perceived agency the entity was measured to have. Not unexpectedly, the humans (teacher and musician with θs of 2.6 and 1.7 respectively) have the highest rated perceived agency while the most repetitive robots have the least (palletizer with θ of -1.0). Table 4 also shows each entity calculated outfit. All entities are within the acceptable outfit ranges from 0.5 to 1.5 except for the video of the "musician."

We can also examine rater latent values α and outfits. Space considerations prevent us from showing a complete table, but their modeled latent value perceived agency, α ranged from 1.93 to -1.80. Of the 186 raters, 26 (14%) had an outfit that was > 1.5. While this number is a bit higher than recommended, it is not too unexpected with online

Entity	θ (SE) Orig	Outfit Orig	θ (SE) Revised	Outfit Revised
Teacher	2.58 (.06)	1.43	2.65 (.06)	1.48
Musician	1.67 (.04)	**1.75**		
Robot Secrets Revealed	1.13 (.03)	1.34	1.17 (.03)	1.40
Bargaining1	.96 (.03)	1.07	.99 (.03)	1.12
Service	.82 (.03)	.96	.85 (.03)	.98
Bargaining2	.73 (.03)	1.03	.75 (.03)	1.11
Cheating	.26 (.03)	.87	.26 (.03)	.90
Firefighting	-.11 (.03)	.79	-.12 (.03)	.81
Line	-.19 (.03)	.87	-.20 (.03)	.90
Soccer	-.48 (.03)	.81	-.50 (.03)	.83
Feeder	-.70 (.03)	.92	-.73 (.03)	.95
Industrial	-.72 (.03)	1.03	-.74 (.03)	1.05
Dishes	-.79 (.03)	.86	-.81 (.03)	.89
Palletizer	-1.00 (.03)	.96	-1.03 (.03)	.96

Table 4. Entity θs and outfits. Original are results with Musician; Revised are results after Musician was removed from the analysis.

participants. The reliability of the rater metrics was .95 which is excellent. Overall, the vast majority of the participants were well modeled by the Rasch analysis.

There are several options to deal with high outfitting entities, items, or raters. Raters that have high outfits can be removed (trimmed), but because Rasch enforces a normal distribution, other raters are likely to become tails. It is also possible to remove some selected scores for raters that had high outfits, but that approach seemed unnecessary given the overall reliability of the data. Since we collected a large number of raters and the impact of individuals is relatively minor (verified by removing the highest 10% outfitting raters), we kept all the raters who passed the attention check.

After observing that "musician" had a high outfit, we examined the comments that raters made on that specific video. Musician was a video of a young woman playing a short concert using a trumpet and a flugelhorn (4-way splitscreen). A bit to our surprise, many raters thought that the video showed a robot playing a musical instrument, showing a fundamental confusion of the video. Re-running the Rasch analysis without the musician showed that all entities and all items were within acceptable ranges (.5 - 1.5); revised results are shown in tables 3 and 4 under the "Revised" headings.

4.3 Discussion

Experiment 1 successfully developed a scale to measure perceived agency. Items were based on our definition and a wide range of entities were rated by over 180 people. We found that the scale was unidimensional and had excellent reliability and separation for each of the facets (entities, items, raters).

There was one item and one entity that showed a moderate outfit; examination of rater comments suggested that one of the entity videos was mis-interpreted so it was removed from the analysis. After removing the high outfitting video, a Rasch analysis confirmed that all items and entities were within acceptable ranges.

4.3.1 Scale Usage. After a scale is created, most researchers will apply it by averaging all the items together for a single value which is then analyzed using traditional statistics (t-test, ANOVAs, linear regression, etc.). Of course running parametric statistics on nominal or ordinal data is not recommended because it loses its inherent meaning (e.g., the difference between 1 and 2 is not necessarily the same as the difference between 4 and 5). Averaging is used because it is easy and because there is usually a monotonic relationship between the raw items and the latent dimension the researchers are trying to measure. Rasch analysis converts the individual scale items into an interval measurement scale, which then can be used in parametric statistics. In our case, we are interested in the amount of perceived agency an entity is judged to have, which is θ in equation 1 and shown in table 4.

As equation 1 shows, calculating the Rasch measure requires knowing the values for each item's β and each person's α. We can calculate an individual's α by asking them to rate known entities, which use as calibration videos. These calibration videos will allow us to determine an individual's α using the algorithm described in [50]. Then we can use each rater's α (predisposition) using the calibration videos, each item's β (item difficulty from Table 3) and their item scores for each entity to determine the amount of perceived agency that a novel entity has according to that rater [50].

Experiment 2 will examine how well the scale constructed in experiment 1 can predict the perceived agency of novel entities.

5 EXPERIMENT 2: SCALE EVALUATION

The goal of experiment 2 was to evaluate the scale constructed in experiment 1 and compare it to other survey methods of measuring perceived agency. As suggested in the introduction, there are two other existing survey approaches that researchers have used to measure perceived agency. The most common is the work by Gray et al. [31]; these items or a subset of these items have been used by others to measure or explore perceived agency [33, 56, 84, 90]. A greatly reduced set of items to measure perceived agency was created by Korman et al. [45]. The items generated by Korman were not psychometrically validated but represents one of the typical [33, 48, 68] approaches used to measure a latent construct: scour the literature and create a "reasonable subset" of items.

Finally, both the averaged raw items and the calibrated items from experiment 1 will be used. There will be four measures of perceived agency that experiment 2 will evaluate on novel entities: (1) the agency dimension from [31]; (2) the agency items from [45]; (3) the average of all 13 items from experiment 1; and (4) the logit scale from experiment 1.

5.1 Method

All studies, including this one were approved by the NRL IRB. All participants consented to participate.

5.1.1 Participants. A monte carlo simulation based on experiment 1 effect sizes showed that 70 participants were needed in order to have an 80% chance of showing a significant ordinal relationship between different entities. 75 participants were recruited through Cloud Research and paid $12 for participation in the study. 10 participants were removed because they missed an attention check ("has a face") leaving 65 participants. The average age of participants was 39 (SD=9) years old. 32 participants were women, 32 participants were men, and 1 participant was unreported. The study took 47 minutes on average. No participants took part in experiment 1.

5.1.2 Materials (Videos). 7 new videos were selected and collected using similar methods as experiment 1. None of the videos in experiment 2 were used in experiment 1.

Table 5 provides a label, a brief description, the morphology of the entity, and a citation of the source. The citation of each video is either a YouTube location or a paper or website describing the video.

5.1.3 Materials (Survey Items). There were three sets of items. One set was developed in experiment 1 and shown in Table 3; these are the PA items. Another set of items came directly from the agency dimension of Gray et al. [31]; these are the GGW items and are shown in Table 6. Finally, the items from Korman et al. [45] and Frazier et al. (under review)

Table 5. Description of videos used in experiment 2.

Label	Entity actions	Morphology	Source
Welding	Welding metal	industrial arm	[85]
TaiChi	Balancing and movement	Humanoid	[62]
Pouring	Pushes cart, unscrews thermos, pours juice and gives it to human	Humanoid	[1]
Robot Secrets Revealed '09	Magician tricking robot	Humanoid	[34]
Bargaining3	Human bargaining with AI agent	Humanoid character	[30]
Punished	Robot put in closet unwillingly	Humanoid	[41, 82]
Professor	Teaching computer science	Human	[88]

[24] are the Korman items and are shown in Table 6. The PA and GGW items used a Likert scale range of 1-5 while the Korman items used a Likert scale range of 1-7.

Table 6. Items and sources used in experiment 2

Item	Source
is capable of conveying thoughts or feelings to others	GGW
is capable of understanding how others are feeling	GGW
is capable of remembering things	GGW
is capable of telling right from wrong and trying to do the right thing	GGW
is capable of making plans and working toward goals	GGW
is capable of thinking	GGW
is capable of exercising self-restraint over desires emotions or impulses	GGW
Did they perform their behavior intentionally?	Korman
Were they aware of engaging in their behavior?	Korman
Did they want to perform their behavior?	Korman

5.1.4 Procedure. The procedure for experiment 2 was identical to the procedure from experiment 1 except for three differences. First, because there were three different scales, we kept each set of survey items together, but the order of each block was randomly determined with a Latin square design.

The second difference from experiment 1 was that after all the videos had been watched and all the items were answered for each video, a ranking screen was displayed. Participants were provided the above definition of perceived agency and asked to rank all the videos from least to most by dragging a thumbnail of each video to the desired rank. They were able to watch any video again if they desired. When this task was completed, they pushed a submit button.

The third difference from experiment 1 was that participants performed a calibration task for three of the entity videos from experiment 1 by answering the PA items from the Table 2. The calibration videos selected were "service" ($\theta = .85$), "cheating" ($\theta = .26$), and "feeder" ($\theta = -.73$): these were selected because they covered a range of perceived agency while not being at the extremes, though in theory, any number of the original videos could be used. The data from the calibration videos was used only for the PA-R scale and did not impact any of the other scales since it was at the end of the experiment.

5.2 Results

5.2.1 Calculating scale values. For the GGW, Korman, and PA scales, the respective items were averaged to give a single score for each rater for each entity. Because of the way Rasch calculates the logit score, only total measures are calculated; reliability can not be calculated for the Rasch measure. However, it is possible to calculate reliability for the raw PA scales; reliability was calculated using α and ω_{total} from the psych package [70].

ω_{total} was .96; Cronbach's α was .95 for PA. ω_{total} was .90; Cronbach's α was .90 for Korman. ω_{total} was .96; Cronbach's α was .95 for GGW.

For the PA-R (Perceived Agency Rasch) measure, the calibration videos were used to calculate each rater's α, or (pre)disposition to perceived agency. Each rater's α was then used with their item ratings to calculate a logit value on an interval scale for each entity [50].

5.2.2 Comparing scales to each other. It is traditional when generating and comparing scales to show the correlations between the different scales. We expect the scales to have moderate to high correlations to each other, since they all attempt to measure the same underlying construct. Indeed, as Table 7 suggests, all the correlations are moderate to high.

Table 7. Correlation matrix between PA, PA-R, Korman, and GGW. [45] and GGW [31] were developed in their respective sources. PA is the average of the Perceived Agency scale developed in experiment 1 while PA-R uses the weights from the Rasch scale developed in experiment 1.

	PA	Korman	GGW
Korman	0.75		
GGW	0.91	0.74	
PA-R	0.84	0.55	0.76

5.2.3 Comparing each scale to empirical ranking. Our overall goal was to determine which scale best predicts the rank ordering of the entities that were ranked from least to most perceived agency by participants. An ordinal regression is the most appropriate analysis for ordered data. An ordinal regression uses an ordinal outcome variable (e.g., rank orderings) while the predictors can be of any type (categorical, ordinal, interval, etc.). Four different ordinal regression models were created, one for each scale.

Figure 1 shows a graphical representation of the ordinal model fits and the empirical data. Model fits were calculated using each rater's scores for each scale to predict a model rank for each entity using the respective ordinal regression model

There are several aspects of figure 1 that should be highlighted. First, notice that all four models fit the empirical data quite well. Even the Korman model [45], which was not constructed or validated in a psychometrically strong manner fits the overall pattern well. The GGW model [31], which was based on a PCA of an agency dimension and is currently the most common method of measuring perceived agency does a very good job of capturing the trends in ordering, though it seems to have the most difficulty in the middle range (i.e., RSR1 and Bargaining are nearly identical in model scores, but quite different empirically). The PA model that consists of the average of items from experiment 1 does an excellent job of predicting the empirical ordering. The PA-R model that was calculated using the the items and three calibration videos from experiment 1 to convert into a logit score using Equation 1 also shows an excellent fit to the empirical data. The difference between the PA and PA-R model is relatively small.

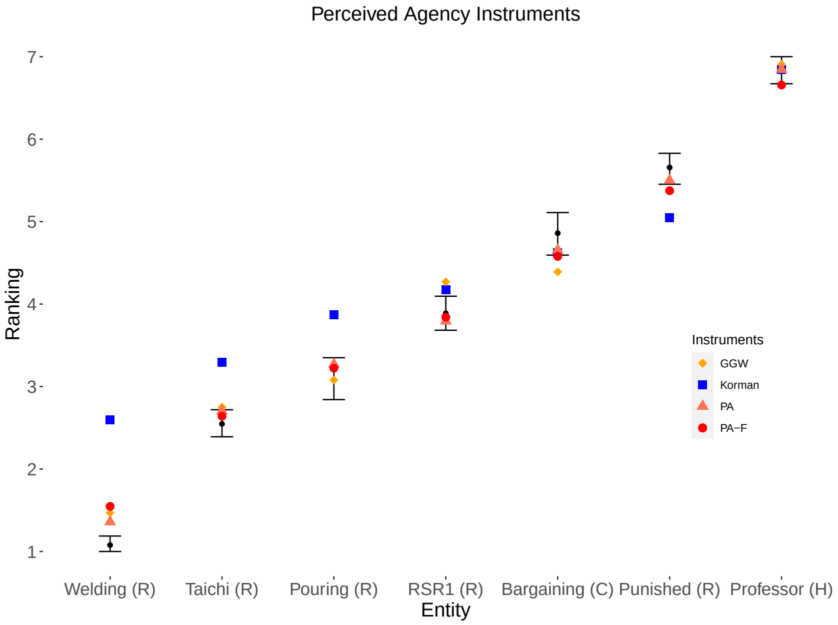

Fig. 1. Results of all models and empirical ranking. (R) is a robot entity; (C) is a character entity; and (H) is a human entity. Black circles are empirical data with 95% CI; model fits are ordinal fits based on each individual model.

We can empirically compare each of the models shown in Figure 1 to determine which is the best predictor of rater rank orderings.

Table 8. Ordinal regression model summaries. Korman [45] and GGW [31] were developed in their respective sources. PA is the average of the Perceived Agency scale developed in experiment 1 while PA-R uses the weights from the Rasch scale developed in experiment 1.

Model	Nagelkerke pseudo R^2	AIC
Korman	.26	1619
GGW	.56	1399
PA	.62	1329
PA-R	.70	1240

Most importantly, all four models are significantly better than chance ($p < 0.05$). We can evaluate how well each model fits the data by using the Akaike Information Criterion (AIC); a lower relative score is better. The AIC statistics derived from the ordinal regression model fits to the data are shown in Table 8. These AIC scores show that the GGW model is significantly preferred over the Korman model, the PA model is significantly preferred over the GGW model, and the PA-R model is significantly preferred over the PA model [89].

5.3 Discussion, Experiment 2

Experiment 2 collected data from a new set of raters on a novel group of seven entities that consisted of a large range of perceived agency. These raters answered items from three different surveys on perceived agency and the results of each of those scales was compared to raters' ranking of the entities.

The ordering of the different entities by the PA and PA-R scale presents a nuanced result of perceived agency. Li et al. (2022) [48] found that humans were rated as having more perceived agency than robots, but our results suggest that not all robots have the same amount of perceived agency. It is not impossible that some robots could have more perceived agency than some humans, and our PA-R scale has the potential to show these more nuanced differences.

Experiment 2 found that all evaluated surveys were acceptable measures of perceived agency and all better than chance. However, the two best surveys were those developed in experiment 1: PA and PA-R. Both PA and PA-R were substantially and significantly better than the other two methods of measuring perceived agency.

5.4 Rasch analysis

Rasch analysis allowed us to construct a measure of perceived agency where all three important facets (entities, items, and raters) were on the same logit scale. Critically, this allows us to examine the hierarchical order of the items: the item βs are estimates of how difficult it is for a rater to agree to each item. This hierarchy allows us to make some important inferences about how people conceptualize perceived agency. First, the items that are most likely for raters to rank highly focus on goals – "acts with purpose" and "has goals:" this is not surprising since an entity without goals can hardly have any perceived agency. In contrast, the items that are the most difficult for raters to rank highly are integrative items (the two scenarios) and emotional items ("can show emotions to other people" and "can change their behavior based on how people treat them." The integrative items highlight that raters will think an entity has a high degree of perceived agency when that entity apparently behaves according to their thoughts, feelings, and not purely responding to the environment. Also, when an entity responds based on their apparent internal feelings, it is more likely to be rated highly on perceived agency. This analysis suggests that robots and other entities that seem to behave according to their internal feelings will be likely to be perceived as having agency.

For the remainder of this paper, we will use PA-R.

6 EXPERIMENT 3: SCALE VALIDATION

Previous researchers have suggested that there is a link between morality and agency [7, 31, 32, 83]. One of the implications of this hypothesis is that entities that have more agency should be protected from harm [7, 31]. We adapted a study from [83] to explore the relationship between perceived agency and harm as well as to validate our measure of perceived agency.

Our hypothesis was that participants should want entities with higher perceived agency to be kept from harm.

6.1 Method

All studies, including this one were approved by the NRL IRB. All participants consented to participate.

6.1.1 Participants. We used the pwr package [13] in R to conduct a power analysis for a correlational study. Our goal was to obtain .8 power to detect a medium sized effect (r=.3) at 0.05 α error probability, so 84 participants were required for this study.

92 participants were recruited through Cloud Research and paid $3.75 for participation in the study. 1 participant was removed because they missed an attention check, leaving 91 participants.

The average age of participants was 41 (SD=12) years old. 43 participants were women, 47 participants were men, and 1 participant was unreported. The study took 16 minutes on average. No participants took part in experiment 1 or 2.

6.1.2 Materials. Eight different images, two instances of four classes, were selected. The four classes were human (images of a man and a woman), dog (images of two dogs); robot (images of a nao robot with a high human likeness and a homemate with a low human likeness [65]), and artwork (pictures of a painting and blown glass).

The scenario was "You are walking down the street and you see an office building across the street from you catch on fire. You call the fire department but rush in to see what you can do to help. You enter and see a single room with [randomly presented] a dog, a robot, a person, and artwork. None of them can move on their own; you will need to carry them outside. Unfortunately, you can only move one at a time. Please drag each item in the order you would move them to safety. You realize that the fire is getting worse."

6.1.3 Procedure. Participants viewed one image of each category (person, dog, robot, artwork) in a random order and answered the perceived agency scale for each entity.

Next participants were given the scenario above and dragged each image in the order they would save it. After they completed dragging each image, they received a message saying, "Congratulations! You managed to save everyone!"

Finally, as in experiment 2, participants saw three calibration videos and answered the perceived agency scale for each calibration video.

6.2 Results

6.2.1 Calculating scale values. Logit scale values for each image were calculated the same way as in experiment 2. The reliability of the perceived agency scale was quite high as well; $\omega_{total} = .98; \alpha = .97$.

6.2.2 Comparing each scale to empirical ranking. Our overall goal was to determine whether there was a relationship between perceived agency and willingness to save an entity or object. Specifically, the higher an entity's perceived agency is, the more likely the entity is to be saved. Statistically, this will be expressed as a negative correlation: higher perceived agency should be negatively correlated with a lower number (i.e., 1 is the first entity/object to save). A simple uncorrected correlation between the saved order of entities and the logit of perceived agency shows there was a strong negative correlation, $r(362) = -.67, p < 0.001$. While this analysis is not technically correct (i.e., it does not take the ordinal variable into account, nor does it take possible inter-dependencies between participants or entities into account), it does give an understandable metric of the relationship.

We can perform a more sophisticated statistical analysis using an ordinal mixed model. Ordinal regression allows us to use an ordinal dependent variable (ordinal data is assumed to violate normality assumptions because the distance between numbers is not metric). A mixed model allows us to take into account correlations between participant or entity ratings.

We can also examine the impact of whether an entity is alive accounts for the negative correlation above; it could be that participants will simply want to save entities that are alive (people, dogs) over inorganic objects (robots, artwork).

We used an ordinal mixed model to analyze the effects of the perceived agency scale and whether the entity was alive on the order of the entities ranked to save considering random variation across participants and images. The analysis was performed using the R package ordinal [15]. We included the entity-saved order (ordinal 1 - 4) as the dependent variable and two independent variables: the logit scores calculated from the perceived agency scale, and whether the object was alive or inorganic (binary). We included as random-effects factors of participants and the stimulus on the intercept.

Unsurprisingly, we found that participants wanted to save entities that were alive sooner than entities that were not, $\beta = -8.00, z = -6.7, p < 0.001$. In addition, and consistent with our hypothesis, the greater an entity's perceived agency, the higher the likelihood to save the entity, $\beta = -.26, z = -2.8, p = 0.005$.

6.3 Discussion, Experiment 3

Experiment 3 examined the hypothesis that when an entity has more perceived agency, people will want it to come to harm less, a component of moral reasoning. We found that an entity's perceived agency did impact the order that it would be saved from destruction. This result goes beyond a simple "save entities that are alive first" heuristic; perceived agency had an effect even after statistically removing the "alive" component.

Experiment 3 also provided construct validity for the PA-R scale: it examined a link between perceived agency and morality that had been suggested in the literature and successfully supported that relationship.

Participants in experiment 3 also used images (not videos) to rate the perceived agency scale. The fact that reliability was excellent suggests that the scale can be used on a variety of different entities and stimulus types.

7 GENERAL DISCUSSION

The goal of this research report was to generate a scale to reliably measure perceived agency and use that scale in a predictive, productive manner. In order to accomplish this goal, we began with a definition of perceived agency. From our definition, we constructed a set of items that were based on each aspect of the definition of perceived agency; this is in contrast to some of the more bottom up approaches (e.g., [31]). Experiment 1 used a Rasch analysis and showed that the scale items were well fitting and that the overall scale had high reliability across all three facets (entities, items, and raters). Experiment 2 used the scale developed in experiment 1 along with two other scales that have been used to measure perceived agency; experiment 2 showed that the scale developed in experiment 1 better captured empirical data than two other current measures of perceived agency.

The PA scale has been developed and tested on a wide variety of entities: videos of humans (3), videos of robots of dramatically different morphologies (15), and videos of AI characters (3). There were also static images of people (2), animals (2), robots (2), and artwork (2). Note that while the majority of entities were humanoid, we also included non-humanoid animals (dogs) and robotic arms and industrial robots and robots with wildly different humanoid features (e.g., wheels, no ears, large eyes, etc.). The successful usage of our scale across this range of entities is encouraging for other entity types.

We should note that these experiments have at least two possible concerns. First, in order to capture a wide range of morphologies, we used videos instead of in-person interactions or observations. Second, the videos were relatively short – less than 3 minutes – and longer interactions may impact the results. We believe, however, that the strength of our approach will overcome these possible weaknesses.

One benefit of measuring how people perceive agency is that we can examine previous work in HRI with our new understanding. We want to emphasize that the success of the created measures supports our definition of perceived agency. We can also provide insight to one of the most influential pieces of work on perceived agency in HRI: Short et al.'s [77] research, who showed that a robot that cheated had more perceived agency than a robot that did not cheat. First, all the conditions had "easy" cues of perceived agency: they had goals, could communicate with others, and because they could communicate and move around, they could perform many different types of tasks and could do well in other environments. However, the cheating robot, in order to cheat, needed to treat others as if they had a mind (thoughts), create novel goals (thoughts), wanted to perform the cheating action (feelings), and could adapt to different situations (losing; environment). These differences are subtle, but note that they also came from each of the three definitional components.

It is our hope that this scale of perceived agency will enable other researchers to accurately measure perceived agency, improve our understanding of how people conceptualize robots' minds, and build robots that have different levels of perceived agency.

ACKNOWLEDGMENTS

The authors thank Anthony Harrison, Sangeet Khemlani, William Adams, Magda Bugajska, Branden Bio, Ed Lawson, and David Porfirio for comments on this research and Mike Linacre and Patrick McKnight for Rasch help.

This work was supported in part by ONR and OSD to GT. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the US Navy.

REFERENCES

[1] 2011. YouTube/Honda unveils all-new ASIMO humanoid robot. https://www.youtube.com/watch?v=1V9XUMCPGF8

[2] 2020. palletizing and box stapling. https://www.youtube.com/watch?v=7vTS67n7wk0

[3] Frances Abell, Frances Happe, and Uta Frith. 2000. Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. *Cognitive Development* 15, 1 (2000), 1–16.

[4] Peter M Asaro. 2007. Robots and responsibility from a legal perspective. *Proc. IEEE* 4, 14 (2007), 20–24.

[5] Jaime Banks. 2019. A perceived moral agency scale: development and validation of a metric for humans and social machines. *Computers in Human Behavior* 90 (2019), 363–371.

[6] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics* 1, 1 (2009), 71–81.

[7] Brock Bastian, Simon M Laham, Sam Wilson, Nick Haslam, and Peter Koval. 2011. Blaming, praising, and protecting our humanity: The implications of everyday dehumanization for judgments of moral status. *British Journal of Social Psychology* 50, 3 (2011), 469–483.

[8] Godfred O Boateng, Torsten B Neilands, Edward A Frongillo, Hugo R Melgar-Quiñonez, and Sera L Young. 2018. Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Frontiers in public health* 6 (2018), 149.

[9] T Bond and C Fox. 2001. Applying the Rasch model. Mahwah, NJ: L.

[10] Selmer Bringsjord. 2008. Ethical robots: the future can heed us. *Ai & Society* 22, 4 (2008), 539–550.

[11] Colleen M Carpinella, Alisa B Wyman, Michael A Perez, and Steven J Stroessner. 2017. The robotic social attributes scale (ROSAS) development and validation. In *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction*. 254–262.

[12] Fulvia Castelli, Francesca Happé, Uta Frith, and Chris Frith. 2000. Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns. *Neuroimage* 12, 3 (2000), 314–325.

[13] Stephane Champely, Claus Ekstrom, Peter Dalgaard, Jeffrey Gill, Stephan Weibelzahl, Aditya Anandkumar, Clay Ford, Robert Volcic, Helios De Rosario, and Maintainer Helios De Rosario. 2018. Package 'pwr'. *R package version* 1, 2 (2018).

[14] Yeh-Tai Chou and Wen-Chung Wang. 2010. Checking dimensionality in item response models with principal component analysis on standardized residuals. *Educational and Psychological Measurement* 70, 5 (2010), 717–731.

[15] Rune Haubo B Christensen. 2015. Analysis of ordinal data with cumulative link models—estimation with the R-package ordinal. *R-package version* 28 (2015), 406.

[16] Kendon J Conrad, Benjamin D Wright, Patrick McKnight, Miles McFall, Alan Fontana, and Robert Rosenheck. 2004. Comparing traditional and Rasch analyses of the Mississippi PTSD Scale: Revealing limitations of reverse-scored items. *Journal of Applied Measurement* 5, 1 (2004), 15–30.

[17] Jose M Cortina, Zitong Sheng, Sheila K Keener, Kathleen R Keeler, Leah K Grubb, Neal Schmitt, Scott Tonidandel, Karoline M Summerville, Eric D Heggestad, and George C Banks. 2020. From alpha to omega and beyond! A look at the past, present, and (possible) future of psychometric soundness in the Journal of Applied Psychology. *Journal of Applied Psychology* 105, 12 (2020), 1351.

[18] DarmstadtDribblers. 2012. https://www.youtube.com/watch?v=CMKX0gZqggA&t=16s

[19] D Dennett. 1978. Brainstorms: Philosophical essays on mind and psychology: Brainstorms e philosophical essays on mind & psychology.

[20] Winand H Dittrich and Stephen EG Lea. 1994. Visual perception of intentional motion. *Perception* 23, 3 (1994), 253–268.

[21] George Engelhard Jr. 2013. *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge.

[22] Michael T Ewing, Thomas Salzberger, and Rudolf R Sinkovics. 2005. An alternate approach to assessing cross-cultural measurement equivalence in advertising research. *Journal of Advertising* 34, 1 (2005), 17–36.

[23] Gerhard H Fischer and Ivo W Molenaar. 2012. Rasch models: Foundations, recent developments, and applications. (2012).

[24] C. R. Frazier, J. M. McCurry, K. Zish, and J. G. Trafton. under review. Perceived Agency Changes Competency Trust and Integrity Trust in Robots. (under review).

[25] R Michael Furr. 2021. *Psychometrics: an introduction*. SAGE publications.

[26] Tao Gao, Gregory McCarthy, and Brian J Scholl. 2010. The wolfpack effect: Perception of animacy irresistibly influences interactive behavior. *Psychological science* 21, 12 (2010), 1845–1853.

[27] Tao Gao, George E Newman, and Brian J Scholl. 2009. The psychophysics of chasing: A case study in the perception of animacy. *Cognitive psychology* 59, 2 (2009), 154–179.

[28] Aimi Shazwani Ghazali, Jaap Ham, Emilia Barakova, and Panos Markopoulos. 2018. The influence of social cues in persuasive social robots on psychological reactance and compliance. *Computers in Human Behavior* 87 (2018), 58–65.

[29] Aimi S Ghazali, Jaap Ham, Emilia I Barakova, and Panos Markopoulos. 2017. Pardon the rude robot: Social cues diminish reactance to high controlling language. In *2017 26th IEEE international symposium on robot and human interactive communication (RO-MAN)*. IEEE, 411–417.

[30] Jonathan Gratch, David DeVault, Gale M Lucas, and Stacy Marsella. 2015. Negotiation as a challenge problem for virtual humans. In *International Conference on Intelligent Virtual Agents*. Springer, 201–215.

[31] Heather M Gray, Kurt Gray, and Daniel M Wegner. 2007. Dimensions of mind perception. *science* 315, 5812 (2007), 619–619.

[32] Kurt Gray and Daniel M Wegner. 2009. Moral typecasting: divergent perceptions of moral agents and moral patients. *Journal of personality and social psychology* 96, 3 (2009), 505.

[33] Kerstin S Haring, Michael Misha Novitzky, Paul Robinette, Ewart J De Visser, Alan Wagner, and Tom Williams. 2019. The dark side of human-robot interaction: ethical considerations and community guidelines for the field of HRI. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 689–690.

[34] A. M. Harrison, B. R. Fransen, M. Bugajska, and J. G. Trafton. 2009. https://www.youtube.com/watch?v=XsubQhtD6S0

[35] Kazuki Hayashida, Yuki Nishi, Michihiro Osumi, Satoshi Nobusako, and Shu Morioka. 2021. Goal sharing with others modulates the sense of agency and motor accuracy in social contexts. *Plos one* 16, 2 (2021), e0246561.

[36] Fritz Heider and Marianne Simmel. 1944. An experimental study of apparent behavior. *The American journal of psychology* 57, 2 (1944), 243–259.

[37] herbmugface. 2017. https://www.youtube.com/watch?v=deP7Gw5JbTU&t=2s

[38] Laura V Herlant, Rachel M Holladay, and Siddhartha S Srinivasa. 2016. Assistive teleoperation of robot arms via automatic time-optimal mode switching. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 35–42.

[39] Kenneth Einar Himma. 2009. Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology* 11, 1 (2009), 19–29.

[40] Ryan Blake Jackson and Tom Williams. 2021. A Theory of Social Agency for Human-Robot Interaction. *Frontiers in Robotics and AI* (2021), 267.

[41] Peter H Kahn Jr, Takayuki Kanda, Hiroshi Ishiguro, Nathan G Freier, Rachel L Severson, Brian T Gill, Jolina H Ruckert, and Solace Shen. 2012. "Robovie, you'll have to go into the closet now": Children's social and moral relationships with a humanoid robot. *Developmental psychology* 48, 2 (2012), 303.

[42] Sara Kiesler, Aaron Powers, Susan R Fussell, and Cristen Torrey. 2008. Anthropomorphic interactions with a robot and robot–like agent. *Social Cognition* 26, 2 (2008), 169–181.

[43] Ryo Kitagawa, Yuyi Liu, and Takayuki Kanda. 2021. Human-inspired Motion Planning for Omni-Directional Social Robots. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 34–42.

[44] Pramote Komolmal. 2014. https://www.youtube.com/watch?v=DiuFkMkReSs

[45] Joanna Korman, Anthony Harrison, Malcolm McCurry, and Greg Trafton. 2019. Beyond programming: can robots' norm-violating actions elicit mental state attributions?. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 530–531.

[46] Sonya S Kwak, Yunkyung Kim, Eunho Kim, Christine Shin, and Kwangsu Cho. 2013. What makes people empathize with an emotional robot?: The impact of agency and physical embodiment on human empathy for a robot. In *2013 IEEE Ro-man*. IEEE, 180–185.

[47] Min Kyung Lee, Jodi Forlizzi, Sara Kiesler, Paul Rybski, John Antanitis, and Sarun Savetsila. 2012. Personalization in HRI: A longitudinal field experiment. In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 319–326.

[48] Zhenni Li, Leonie Terfurth, Joshua Pepe Woller, and Eva Wiese. 2022. Mind the Machines: Applying Implicit Measures of Mind Perception in Social Robotics. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*. 236–245.

[49] John Linacre. 1994. Sample size and item calibration stability. *Rasch Mes Trans.* 7 (1994), 328.

[50] JM Linacre. 1998. Estimating measures with known polytomous item difficulties. *Rasch Meas Trans* 12 (1998), 638.

[51] John Michael Linacre. 2006. Demarcating category intervals. *Rasch Measurement Transactions* 19, 3 (2006), 341–43.

[52] John Michael Linacre. 2022. Facets. *Computer Program for Many-faceted Rasch Measurement* (2022).

[53] John M Linacre, MH Stone, J William, P Fisher, and L Tesio. 2002. Rasch Measurement. *Rasch Measurement Transactions* 16 (2002).

[54] Alexandru Litoiu, Daniel Ullman, Jason Kim, and Brian Scassellati. 2015. Evidence that robots trigger a cheating detector in humans. In *Proceedings of the tenth annual acm/ieee international conference on human-robot interaction*. 165–172.

[55] Bertram Malle. 2019. How many dimensions of mind perception really are there?. In *Proceedings of CogSci*. 2268–2274.

[56] Bertram F Malle. 2021. What the mind is. *Nature Human Behaviour* (2021), 1–2.

[57] Molly C Martini, Christian A Gonzalez, and Eva Wiese. 2016. Seeing minds in others–Can agents with robotic appearance have human-like preferences? *PloS one* 11, 1 (2016), e0146310.

[58] Eric Martinson, Wallace E Lawson, Samuel Blisard, Anthony M Harrison, and J Greg Trafton. 2012. Fighting fires with human robot teams.. In *IROS*. 2682–2683.

[59] D Betsy McCoach, Robert K Gable, and John P Madura. 2013. *Instrument development in the affective domain*. Vol. 10. Springer.

[60] Quinn McNemar. 1946. Opinion-attitude methodology. *Psychological bulletin* 43, 4 (1946), 289.

[61] Carey K Morewedge, Jesse Preston, and Daniel M Wegner. 2007. Timescale bias in the attribution of mind. *Journal of personality and social psychology* 93, 1 (2007), 1.

[62] Motorward. 2017. https://youtu.be/jJYsOsoBIZU?t=243

[63] Tatsuya Nomura, Takayuki Kanda, Tomohiro Suzuki, and Kennsuke Kato. 2004. Psychology in human-robot communication: An attempt through investigation of negative attitudes and anxiety toward robots. In *RO-MAN 2004. 13th IEEE international workshop on robot and human interactive communication (IEEE catalog No. 04TH8759)*. IEEE, 35–40.

[64] Stephen Nowicki and Marshall P Duke. 1974. A locus of control scale for noncollege as well as college adults. *Journal of Personality Assessment* (1974).

[65] Elizabeth Phillips, Xuan Zhao, Daniel Ullman, and Bertram F Malle. 2018. What is human-like? decomposing robots' human-like appearance using the anthropomorphic robot (abot) database. In *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*. 105–113.

[66] Gilles Raîche. 2005. Critical eigenvalue sizes in standardized residual principal components analysis. *Rasch measurement transactions* 19, 1 (2005), 1012.

[67] rdiankov. 2009. https://www.youtube.com/watch?v=gySlayBF3v4

[68] Samantha Reig, Elizabeth J Carter, Terrence Fong, Jodi Forlizzi, and Aaron Steinfeld. 2021. Flailing, hailing, prevailing: Perceptions of multi-robot failure recovery strategies. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 158–167.

[69] William Revelle. 2022. An introduction to psychometric theory with applications in R.

[70] William Revelle and Maintainer William Revelle. 2015. Package 'psych'. *The comprehensive R archive network* 337, 338 (2015).

[71] William Revelle and Richard E Zinbarg. 2009. Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika* 74 (2009), 145–154.

[72] Julian B Rotter. 1954. Social learning and clinical psychology. (1954).

[73] Perrine Ruby and Jean Decety. 2001. Effect of subjective perspective taking during simulation of action: a PET investigation of agency. *Nature neuroscience* 4, 5 (2001), 546–550.

[74] Peter AM Ruijten, Antal Haans, Jaap Ham, and Cees JH Midden. 2019. Perceived human-likeness of social robots: testing the Rasch model as a method for measuring anthropomorphism. *International Journal of Social Robotics* 11, 3 (2019), 477–494.

[75] Brian J Scholl and Patrice D Tremoulet. 2000. Perceptual causality and animacy. *Trends in cognitive sciences* 4, 8 (2000), 299–309.

[76] Johannes Schultz and Heinrich H Bülthoff. 2013. Parametric animacy percept evoked by a single moving dot mimicking natural stimuli. *Journal of vision* 13, 4 (2013), 15–15.

[77] Elaine Short, Justin Hart, Michelle Vu, and Brian Scassellati. 2010. No fair!! an interaction with a cheating robot. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 219–226.

[78] Jagdip Singh. 2004. Tackling measurement problems with Item Response Theory: Principles, characteristics, and assessment, with an illustrative example. *Journal of Business Research* 57, 2 (2004), 184–208.

[79] Richard M Smith and Kyunghee K Suh. 2003. Rasch fit statistics as a test of the invariance of item parameter estimates. *Journal of applied measurement* 4, 2 (2003), 153–163.

[80] Everett V Smith Jr. 2002. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of applied measurement* 3, 2 (2002), 205–231.

[81] Paul E Spector. 1988. Development of the work locus of control scale. *Journal of occupational psychology* 61, 4 (1988), 335–340.

[82] IEEE Spectrum. 2012. https://www.youtube.com/watch?v=DAiWZO0dz8M

[83] Megan Strait and Matthias Scheutz. 2014. Using functional near infrared spectroscopy to measure moral decision-making: effects of agency, emotional value, and monetary incentive. *Brain-Computer Interfaces* 1, 2 (2014), 137–146.

[84] Tetsushi Tanibe, Takaaki Hashimoto, and Kaori Karasawa. 2017. We perceive a mind in a robot when we help it. *PloS one* 12, 7 (2017), e0180952.

[85] Olympus Technologies. 2017. https://www.youtube.com/watch?v=Oz7TE1Q1rhw

[86] Maia Trafton. 2021. https://www.youtube.com/watch?v=NtWaLJ-N6u0

[87] Daniel Ullman and Bertram F Malle. 2018. What does it mean to trust a robot? Steps toward a multidimensional measure of trust. In *Companion of the 2018 acm/ieee international conference on human-robot interaction*. 263–264.

[88] Harvard University. 2016. https://www.youtube.com/watch?v=0JUN9aDxVmI&t=4273s

[89] Eric-Jan Wagenmakers and Simon Farrell. 2004. AIC model selection using Akaike weights. *Psychonomic bulletin & review* 11, 1 (2004), 192–196.

[90] Kara Weisman, Carol S Dweck, and Ellen M Markman. 2017. Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences* 114, 43 (2017), 11374–11379.

[91] Astrid Weiss and Christoph Bartneck. 2015. Meta analysis of the usage of the godspeed questionnaire series. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 381–388.

[92] Eddie Woo. 2014. https://www.youtube.com/watch?v=EAyI-yiXaek

[93] Benjamin D Wright and Mark H Stone. 1979. Best test design. (1979).

[94] Shannon Yasuda, Devon Doheny, Nicole Salomons, Sarah Strohkorb Sebo, and Brian Scassellati. 2020. Perceived Agency of a Social Norm Violating Robot. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

A EARLIER DATA COLLECTION

Before experiment 1 occurred, we ran 2 previous data collections. The method was similar to experiment 1, though there were some small methodological differences (i.e., we had each participant rate 14 videos instead of splitting them up; participants were encouraged to provide feedback on the clarity of the items); here we show the items that were generated and why they were changed or removed and how we came to the items in experiment 1. These two data collections allowed us to go into experiment 1 with high expectations that they would be good items for measuring perceived agency.

A.1 Data Collection A

Data collection A collected responses from 109 online participants and had 23 items. To generate a broad item pool, we started with previous definitions and previous studies of perceived agency. The items were based on physical similarity and associated affordances [3, 12, 27, 36, 42, 57, 61, 76], emotional expression and recognition [28, 29, 31, 40], self-directed goals [20, 35, 36, 39, 75], interaction with others [28, 36, 40, 47, 73, 77], and having general mental capabilities [31, 40, 90]. There were two primary concerns with the items after running a Rasch analysis. First, many of the items and raters had especially high outfits, suggesting that the raters had difficulty rating the videos in a consistent manner. Second, Rasch analysis shows the probability of item categories being in a different category; for this analysis three of the four category thresholds were within .05 logits, suggesting that either the number of categories was too large or raters could not use the items to consistently differentiate the entities in the videos. We interpreted these results as showing that this set of items was too broad to be measured unidimensionally with consistency. We therefore narrowed and crystallized our definition (see section 2) while removing items that participants found difficult.

A.2 Data Collection B

Data collection B collected responses from 130 online participants and had 14 items. The 14 items came from our definition (see section 2). Data collection B mixed Likert responses with semantic scales which some participants found confusing. Semantic scales were removed for future experiments and/or converted to Likert responses. Experiment 1 was the result of the updated items.

Table 9. Experiment A items and reasons for keeping or rejecting them

Item	Reason for keeping or rejecting
can perceive the environment	perception per se not part of PA
can respond to things in the environment	too vague
can navigate in the environment	navigation per se not part of PA
can manipulate objects in the environment	manipulation per se not part of PA
can tell one object from another object	object identification per se not part of PA
can show emotions	kept for final version with small modification
can show likes OR dislikes	too similar to preferences
can show preferences	confusing so changed to ’wanted to perform these actions’
can recognize emotions	recognizing emotions per se not part of PA
can make plans and work towards a goal	converted to intentionality
can change their own actions in response to others	modified and kept for final
can change their own actions in response to the environment	modified and kept for final
has free will	kept but then later removed
is being directly controlled by another	difficult to understand for non-robot entities
understands how others feel	modified
knows right from wrong	modified
recognizes that their actions have impact on others	modified
communicates with others	communication per se not part of PA
interacts with others	interaction per se not part of PA
treats others as if they were alive	too vague
can remember things	memory is probably a subcomponent of PA but not specific enough so removed
realizes when it has made a mistake	mistake identification / meta-cognition too vague
is able to learn	learning may be a subcomponent of PA but not specific enough so removed

Table 10. Experiment B items and reasons for keeping or rejecting them

Item	Reason for keeping or rejecting
has some understanding of how others think and feel	combined thoughts and feelings so removed
treats others as if they had a mind	kept
can adapt to different situations	kept
knows morally right from morally wrong	morality not inherently part of PA
acted intentionally	modified to make more understandable
wanted to perform these actions	kept
was aware of engaging in their actions	overlapped with ’wanted to perform’
actions were driven entirely by the environment	’entirely’ was too strong; changed to environmental items
acts automatically vs. acts purposefully	broken into environmental items and thoughts/goals items
acts without understanding vs. understands their actions	understanding per se not part of PA
free will	part of consciousness, not necessarily PA
actor scenario	kept
birdhouse scenario	required too much manipulation and similar to other scenarios
dinner scenario	kept