Scientific Reports

https://doi.org/10.1038/s41598-025-29685-w

Article in Press

Measuring transparency in intelligent robots

Received: 2 September 2024

Accepted: 18 November 2025

Published online: 12 December 2025

Cite this article as: Angelopoulos G., Lacroix D., Wullenkord R. *et al.* Measuring transparency in intelligent robots. *Sci Rep* (2025). https://doi.org/10.1038/s41598-025-29685-w

Georgios Angelopoulos, Dimitri Lacroix, Ricarda Wullenkord, Alessandra Rossi, Silvia Rossi & Friederike Eyssel

Authors' Response to Reviews of

Measuring Transparency in Intelligent Robots

Georgios Angelopoulos, Dimitri Lacroix, Ricarda Wullenkord, Alessandra Rossi, Silvia Rossi, Friederike Eyssel

Nature Scientific Reports,

RC: ***Reviewers' Comment***, AR: Authors' Response, ☐ Manuscript Text

Dear Editor and reviewers,

We thank you for your valuable feedback on our manuscript. We appreciate the time and effort put into reviewing our work and providing constructive comments. We have carefully considered this feedback and have revised the manuscript accordingly. In the following, please find our point-by-point responses. We have marked changes to the manuscript in yellow.

Sincerely,

Dimitri Lacroix on behalf of the authors

1. Reviewer #1

We would like to thank the reviewer for their encouraging remarks on the quality and potential of our work. We have carefully revised the manuscript and supplementary materials to address all concerns.

RC: ***1. I am not a fan of the introduction as currently written. First, I would like to get more quickly to the theoretical foundations of the three components explainability, legibility, and predictability (page 2). Second, while I appreciate the authors' attempts to illustrate the construct of transparency on page 1, I believe the example, in an odd way, muddles the picture. Both it, and the discussion in the second paragraph of the introduction, seem to mistake intention (and intentionality, i.e., the what?) for transparency (and, in particular, explainability, i.e., the why? and how?). Instead, I would suggest moving up the discussion in the two paragraphs on page 2 as early as possible, and to clarify the three components early on as well.***

AR: We thank the reviewer for the valuable suggestions. Following Reviewer 1, we have reorganized the introduction so that the theoretical foundations of explainability, legibility, and predictability appear earlier. The example on page 1 has been removed to avoid conflating intentionality with transparency.

RC: ***2. As the authors briefly acknowledge, but then not really discuss, using reverse-coded items, while often recommended, typically results in factor structures that more reliably distinguish between positively and negatively worded items, rather than increase reliability within conceptual factors that contain both negatively and positively phrased items. Therefore, I found it odd that the authors in Stages 1 and 2 did not really discuss the target constructs of the original item pool and that they did not get to name the three factors until very, very late in the paper. As a result, I believe there is a disconnect between the targeted and theory-based four factors initially described, on one hand, and the empirically derived factor structures with three factors, on the other. Pair this with the fact that, at least from my review, it appears as if factor 1 (later labeled "illegibility") is the only one that contains any (and indeed only) negatively worded/reverse-coded items, and one must ask the question whether there is a threat of an error of reification in naming it "illegibility", rather than "negatively phrased items". Maybe the authors can***

discuss this more, or test alternative models via CFA (see also concern D. below).

AR: Prior research has indeed demonstrated that reverse-worded items, such as polar opposites (e.g., “It is difficult for me [...]” instead of “It is easy for me [...]”) or negated items (e.g., “It is not easy for me [...]” instead of “It is easy for me [...]”), may threaten the reliability and validity of a scale[1, 2, 3]. Such items might create “method factors”, with EFAs commonly distinguishing reverse-worded and regular items as two separate factors where one would expect a single one[1, 3, 4]. Moreover, items that include negations appear to be more challenging, difficult to interpret, and can elicit fatigue [2, 3]. Therefore, when we realized during item selection at Stage 2 that Factor 1 mainly included reverse worded items, we had the same concerns as Reviewer 1. However, several aspects pertaining our results convinced us to keep Factor 1 as it is. First, the EFA conducted at Stage 2 did not identify a clear method factor. To illustrate, even after item removal, Factor 1 contained two "regular" items (i.e., “It is clear to me what the robot does” and “I have a clear understanding of how the robot operates in general.”). Furthermore, Factor 1 was not the only factor containing reverse-worded, as Factor 3 comprised two reverse-worded items (i.e., “It is difficult for me to tell what the robot will do next.”, and “The robot’s behavior does not help predict what it will do next.”). This, however, is not sufficient to discard the existence of a method factor due to reverse-worded items. To do so, Reviewer 1’s recommendation to run alternative CFAs is indeed highly relevant and we followed it (see our answer to RC 4 for more details). One of them tested a model distinguishing reverse worded and regular items as two separate factors, and resulted in poor fit indices, which provided more support for the idea that Factor 1 is not a method factor.

Regarding the effect of reverse-worded items on the difficulty , the item difficulty of all the items of the scale (i.e., range between .580 - .710), including Factor 1, was within the range of acceptability (i.e., between .200 - .900) [5, 6, 7, 8] and close to the optimal range (i.e., between .500 and .600, with some authors extending the upper limit up to .750) [7, 8]. As a matter a comparison, the average and range of item difficulty for each factor were as follows:

Factor 1 (Illegibility): $M = .651$, range: $.600 - .710$

Factor 2 (Explainability): $M = .654$, range: $.580 - .690$

Factor 3 (Predictability): $M = .670$, range: $.650 - .690$

The three factors have almost equal difficulty, suggesting that participants’ answers to the items from these different factors do not meaningfully change depending on how the items are phrased. We deem that the inclusion of mostly reverse worded items in Factor 1 did not make it more challenging for participants to answer its items. We added the average and range of item difficulty and discrimination as a note below table S14 of the supplementary material. In the section dedicated to the Item generation, we also mentioned that the response options of our scale were fully labelled, which increases the understandability of the items, reduces cognitive load of respondents, and limitates the risk of misresponse to reverse-worded items [9].

The second reason for us to keep Factor 1 as it is and naming it illegibility, whilst data-driven, has a theoretical rationale. Indeed, we assumed that the items loading on Factor 1 were mostly reverse-worded ones because a perceived lack of legibility is probably a more salient information to evaluate a robots’ transparency than legibility (i.e., observers do not notice when a robot performs actions that comply to social norms or expectations). We provided this theoretical rationale in the discussion.

We acknowledge that these methodological and theoretical points should have been included in the manuscript in the first place. Most importantly, we agree with Reviewer 1 that there was a lack of continuity between our theoretical assumption about the factorial structure of transparency, and the conclusions we drew from our research. The revision of the introduction and the discussion is intended to connect these parts better, while

more details were provided in each experiment to make their results more understandable. In particular, the content of the items per extracted Factor at Stage 1 is commented.

RC: ***3. I would like the authors to comment a little more on the use of a traditional 7-point Likert-type question format. Personally, I am not a fan of the neutral item in 3-, 5-, and 7-point scales, especially for items such as "I feel confident..." What exactly does a "4" = "Neutral" (Neither agree or disagree) mean here? Please comment and discuss whether a 6-point scale might be better.***

AR: We would like to thank Reviewer 1 for raising this too often overlooked aspect of scale development. As we mentioned in the manuscript, we opted for a 7-point scale based on existing recommendations. However, we did not mention our rationale for making a scale with a an odd response format instead of an even response format. Midpoints allow participants to express when their evaluation is neutral (i.e., is both lowly positive and negative) or ambivalent (i.e., a highly positive and negative conflicting evaluation [10, 11, 12]. This is especially important when considering ambivalent respondents, as they tend to react negatively when they are constrained to make a choice [12, 9]. We added this rationale in the last paragraph of Stage 1 section dedicated to the item generation. Furthermore, although midpoints have been identified as potential threats to the reliability of a scale, available empirical evidence for such an effect is inconsistent [13]. Given that the reliability indicators (i.e., Cronbach's alpha, McDonald's Omega) of our scale were very high, we believe the presence of midpoints do not put the reliability of our scale at stake. We included this argument in the results section of Experiment 1.

RC: ***4. Also related to B above. As you had enough data for confirmatory factor analyses across languages, did you compare the fit of the EFA-suggested original structure from Stage 2 to any alternative factor structures (1-factor, positive vs negatively phrased items, theoretically suggested, random, etc.) in any of the languages, or across all? I would like to see comparative fit across different alternative models, rather than only fit of the EFA-suggested model.***

AR: As suggested, we conducted additional CFAs testing. Three models were tested: The first one was dedicated to test the presence of a method factor by gathering all reverse worded items in a factor, and all regular items in another factor. This two-factor CFA showed poor fit indices in all three languages. The second CFA tested whether all items would not be better explained by a single factor, namely transparency. This 1-factor CFA also resulted in poor fit indices in all three languages. Finally, a 4-factor model was tested, based on our initial theoretical assumption that transparency is composed of legibility, explainability, predictability, and meta-understanding. Fit indices of this 4-factor CFA ranged from acceptable to excellent. Comparatively, the three factor model however showed the best fit indices (while being more parcimonious). The results of these alternative CFAs can be consulted in Tables S21-S23 of the supplementary material

RC: ***5. Very minor concern: Somehow "consensual definition" both sounds weird to me, and it is also overused. How about "consensus definition" instead, and maybe some variety like "commonly accepted definition"?***

AR: We thank the reviewer for pointing this out. We replaced "consensual definition" with "commonly accepted definition".

2. Reviewer #3

We thank Reviewer #3 for the very positive assessment of our methodology, data presentation, and appendices. We also appreciate the constructive feedback on areas of improvement.

RC: ***1. I would suggest reorganizing your introduction: you talked a lot about the importance of transparency***

in HRI or even human-human interaction while avoiding a clear definition. Then, later, you discuss why it is difficult and all the different definitions, but I did not clearly see your own take. I have several candidates from the introduction, but I recommend it is stated early and explicitly to help readers understand early discussion. Rather than saying "other research suggests a multifaceted approach ..." you can say, based on work, we take a multifacted initial definition that combines (factors). To my understanding, this soft definition was likely taken as the authors likely understood the definition may change based on factor loadings etc, but given that they had validation scenarios in mind that could differentiate transparency (high/low), they clearly had at least some example scenarios that shared a common definition. In short, I understand it's complex, but help the readers better early on understand what you mean by transparency.

AR: We thank the reviewer for this helpful suggestion. In the revised introduction, we now provide an explicit working definition of transparency at the outset, grounded in prior literature but clearly framed as our initial perspective.

RC: ***2. Relatedly, the other major issue I had was about impact and why/where I or other practitioners should use this scale. And this relates to an issue I had: I'm unclear why MDMT and UTAUT were selected as tests for correlation. I understand trust is related theoretically, but it is still a different application. For example, if I trust someone or a robot, I may not need it to be transparent, or vice versa. Or, if they are considered overlapping concepts, this is not clearly explained, or what these scales are chosen over other trust questionnaires (or other related concepts). Further, after reading the results of Stage 3, I wonder, if they correlate so strongly with MDMT and UTAUT, do we need this scale? Are all 3 not measuring some latent factor themselves? I was hoping to see that MDMT or UTAUT failed to measure or explain results as well as the TOROS in a study designed specifically to differentiate transparency. Perhaps I misread the results, but I do think the choices of these to check correlation/validation should be better justified, and there should be some discussion on both theoretically how these concepts differ and if these questionnaires truly measure different things, if they correlate so strongly.***

AR: We thank the reviewer for bringing up this critical point. As no validated scale for perceived transparency was available to assess convergent validity with other measures, we selected two constructs that transparency is supposed to determine: Acceptance and trust. We agree with Reviewer 3 that trust is distinct, but related to transparency. Accordingly, we made the theoretical distinction between both constructs clearer in the introduction. However, prior work showed that manipulating robot transparency has a positive effect on acceptance and trust towards robots [14, 15, 16, 17]. Therefore, we expected our measure of perceived transparency to correlate with MDMT and UTAUT not because they are assumed to be the same constructs, but because one (transparency) determines the others.

It must be noted that correlations do not necessarily suggest a construct overlap [18]. In addition, trust and acceptance do not explain our results, as they were used as dependent variables themselves, and not included as predictors in any analysis (e.g., regression models). They were indeed sensitive to our manipulation of transparency in both experiments, which is consistent with prior research [14, 15, 16, 17]. However, our experimental manipulation of transparency in both studies had the most significant and large effects on the different dimensions of transparency, as measured with TOROS, which shows that our measure of transparency, while being a much more direct measure than relying on measures of trust or acceptance, is also more sensitive to induced changes of perceived transparency than measures of trust and acceptance.

We believe that this is not a misreading of the results, but rather a lack of clarity on our part. Therefore, we edited the results and the discussion of both Experiments 1 and 2 to clarify what the results imply. In addition, more details about the use of the scale in further research is provided in the discussion, so readers have a better idea of how they can benefit from using it, and what the best use of the scale is.

RC: ***3. I only recommend considering in-person validation, even briefly, as the paper uses only online, "bystander observation" style only video surveys. I don't think this significantly weakens the paper, but I do think it would be much stronger with such a validation.***

AR: We agree with the reviewer and have added a paragraph in the Limitations section highlighting the absence of in-person validation. We emphasize that this was a pragmatic choice to ensure large, cross-linguistic samples.

RC: ***4. While cross-cultural validation is interesting, cultural factors potentially affecting transparency (what means transparency across cultures, cultural views towards and values of transparency) could be better explored. Outside of convenience, I would be curious if there was any specific reasons why these 3 languages were picked. On a related point, the Acknowledgements and contributions state both translators and authors translated the scale into the respective languages. While it makes sense for the authors to help with nuances or intent from the research perspective that may not be clear to translators, I'm curious about this process, if their was a difference between languages, and if the authors believe these processes or translations themselves were the reason behind the partial invariance of the results between languages.***

AR: We agree that cultural factors affecting transparency would be important to investigate in further research. To that end, we elaborated on future attemps to translate and validate the scale in other languages in the discussion.

Regarding the targeted languages for the validation (i.e., English, German, and Italian), they were indeed partly motivated by the fact the authors belong either to a German or an Italian institution, hence the potential need to measure transparency in German or Italian language. Moreover, English speaking, German speaking, and Italian speaking countries are among the most prolific countries in terms of research on HRI. We are also currently working on a Greek and a French translations of the scale, but we acknowledge that other languages might be extremely relevant (e.g., Japanese or Korean), and we will progressively extend the translations based on our resources.

Regarding the translation process, one author performed the first round of translation from English to German, and another one from English to Italian (both were native speakers of the targeted language). For each translation, an Italian-speaking and German-speaking naive translator (both native speakers of Italian and German as well) were requested to backtranslate these scales into English language. Adjustments were then collectively made based on observed discrepancies between the original scale and each scale backtranslated in English language. We recruited two independent translators, who were unaware of the research purpose, performing the translation from the original language to the targeted language. Once each translator finished their translation, they discussed the discrepancies between translations with the researchers to create a single one. This translation was then given to another team of independent translators who backtranslated the scale in the original language. Once again a collective discussion was set up to adjust the translations based on the discrepancies between the original scale and its backtranslated version [19]. However, the most critical aspect of scale translation is to have translators be independent from each other, which we complied to [19]. We believe that having authors translating from English to a targeted language and leaving naive translators backtranslating was our best solution to limit potential biases. However, we cannot demonstrate that partial invariance was not due to our translation process.

RC: ***5. Figure 2 has, apparently, a "clear breakpoint after 5 factors". I think the clearest breakpoint is after 1, perhaps after 4. I'm unclear why 5 is supposedly so clear - how is this chosen? I agree from the data 5 factors is lower than 4, but so is 9 compared to 5. Perhaps the decrease from 3 to 5, while visually detectable, is not actually that useful in accuracy? Some theoretical or logical justification should be given outside of stating it being "clear" as it may not always be clear to readers. This is especially confusing as you justify early on the ideas of Legibility, predictability, and explainability, and then add in meta-understanding***

which disappears conceptually later on. It turns out you ended with 3 factors later anyways, so maybe it was not so clear after all, and the discussion of meta-understanding is more of a future work/discussion point rather than an introduction point

AR: We thank Reviewer 3 for pointing out this mistake, as the breakpoint is at 2 factors, and the curve flattens at 4 to 5 factors. The latter aspect of the graph guided our decision. We corrected this part of our paper and added further information regarding how we chose the initial number of factors to retain. Since this analysis was the initial step of item removal, we opted for 5 factors as a way to avoid both the early removal of potentially meaningful factors (i.e., by opting for a 1- or 2-factor model) and a model overfit (i.e., by solely relying on eigenvalues above 1.0, which suggested 7 factors).

The disappearance of meta-understanding is discussed in the results of Experiment 1. Specifically, we commented the content of Factor 1, that encompassed items supposed to measure both legibility and meta-understanding, suggesting that human observers do not deem a robot's action legible because they can understand the action itself, but rather because they can make inferences regarding the robot's internal states (e.g., goal, functioning). We followed Reviewer 3's advice to remove meta-understanding from the introduction. As this concept was integral to the initial development of the 64 items to assess perceived transparency of robots' behavior (and thus part of the preregistration), we briefly mentioned it in the Item Generation section. Its disappearance from the scale after at the end of the item removal procedure is also commented in Stage 2 section.

RC: ***6. I appreciated the introduction on Item Response Theory as it was new to me and the paper cited for it is quite interesting. While I was familiar with other methods used here, like EFA, I think a rough overview of methods (at least explaining their purpose in one or two sentences) could be very useful for less experienced researchers and general readability.***

AR: We appreciate Reviewer 3's consideration for our theoretical background regarding scale development and testing. Given the space constraints, an overview of methods would have been too dense and to some extent out of the scope of our research. However, we followed the recommendation to include a rough explanation of both exploratory and confirmatory factor analyses (in the introduction) and Item Response Theory (in Stage2 section). We also added references regarding these methods so the less experienced readers can find relevant and detailed information to understand our work and use these approaches in their own research.

RC: ***7. While more of a typesetting issue, I was confused as I thought Table 2 was related to Stage 3 of the development. Consider rephrasing the captions to better describe what part of the process they belong to, so if the tables or figures are distributed in other pages or sections, readers can more quickly understand what they are for.***

AR: We followed the reviewer's recommendation to indicate the experiment to which a table or figure is related in the captions.

RC: ***8. I'm curious why the authors ended up naming factor one negatively, Illegibility instead of Legibility. Is it that the questions themselves focused on this negative view? I'm curious if there is some reason or meaning in the fact that participants responded more to questions in the negative (I cannot understand what this robot is thinking) vs positive (I can understand what this robot is thinking).***

AR: We thank the reviewer for this thoughtful comment. The decision to label Factor 1 as Illegibility rather than Legibility was both empirically and theoretically motivated: Empirically, the items loading on this factor were predominantly phrased as negations. This reflects the way participants seemed to engage more with limitations in their understanding rather than with its presence. Theoretically, we interpret this as suggesting that a lack of legibility may be more salient and accessible for participants when evaluating a robot's transparency. People

notice breakdowns or absences of clarity more readily than smooth and successful communication. These considerations have been included in the discussion, with relevant references.

Reviewer 1 raised the concern that the presence of mostly reverse-worded items in Factor 1 at the end of Stage 1 could be an artefact, as factor analyses tend to distinguish regular items and reverse-worded items as two different factors, even when these items are supposed to capture the same construct [1, 2, 3]. Following Reviewer's 1 suggestion, a 2-factor CFA distinguishing regular items and reverse-worded items was conducted, which resulted in poor fit-indices. Alongside other alternative CFAs, the results provided support for our three-factor model being the best factor structure to explain the data of TOROS.

References

[1] X. Zhang, R. Noor, and V. Savalei, "Examining the Effect of Reverse Worded Items on the Factor Structure of the Need for Cognition Scale," *PLoS ONE*, vol. 11, p. e0157795, June 2016.

[2] C. M. Woods, "Careless Responding to Reverse-Worded Items: Implications for Confirmatory Factor Analysis," *Journal of Psychopathology and Behavioral Assessment*, vol. 28, pp. 186–191, Sept. 2006.

[3] E. van Sonderen, R. Sanderman, and J. C. Coyne, "Ineffectiveness of Reverse Wording of Questionnaire Items: Let's Learn from Cows in the Rain," *PLOS ONE*, vol. 8, p. e68967, July 2013.

[4] B. Weijters, H. Baumgartner, and N. Schillewaert, "Reversed item bias: An integrative model," *Psychological Methods*, vol. 18, no. 3, pp. 320–334, 2013. Place: US Publisher: American Psychological Association.

[5] S. K. S. Shanmugam, V. Wong, and M. Rajoo, "Examining the quality of english test items using psychometric and linguistic characteristics among grade six pupils: Examining the quality of test items using psychometric properties in classical test theory," *Malaysian Journal of Learning and Instruction*, vol. 17, no. 2, pp. 63–101, 2020.

[6] P. Mukherjee and S. K. Lahiri, "Analysis of Multiple Choice Questions (MCQs): Item and Test Statistics from an assessment in a medical college of Kolkata, West Bengal," *IOSR Journal of Dental and Medical Sciences*, vol. 14, no. 12, pp. 47–52, 2015.

[7] L. S. Feldt, "The Relationship Between the Distribution of Item Difficulties and Test Reliability," *Applied Measurement in Education*, Jan. 1993. Publisher: Lawrence Erlbaum Associates, Inc.

[8] J. Collins, "Education techniques for lifelong learning: writing multiple-choice questions for continuing medical education activities and self-assessment modules," *Radiographics: A Review Publication of the Radiological Society of North America, Inc*, vol. 26, no. 2, pp. 543–551, 2006.

[9] B. Weijters, E. Cabooter, and N. Schillewaert, "The effect of rating scale format on response styles: The number of response categories and response category labels," *International Journal of Research in Marketing*, vol. 27, pp. 236–247, Sept. 2010.

[10] K. J. Kaplan, "On the ambivalence-indifference problem in attitude theory and measurement: A suggested modification of the semantic differential technique," *Psychological Bulletin*, vol. 77, no. 5, pp. 361–372, 1972. Place: US Publisher: American Psychological Association.

[11] J. G. Stapels and F. Eyssel, "Let's not be indifferent about robots: Neutral ratings on bipolar measures mask ambivalence in attitudes towards robots," *PLOS ONE*, vol. 16, p. e0244697, Jan. 2021. Publisher: Public Library of Science.

[12] M. Gilljam and D. Granbi, "Should we take don't know for an answer?," *Public Opinion Quarterly*, vol. 57, pp. 348–357, Jan. 1993.

[13] S. Y. Y. Chyung, K. Roberts, I. Swanson, and A. Hankinson, "Evidence-Based Survey Design: The Use of a Midpoint on the Likert Scale," *Performance Improvement*, vol. 56, no. 10, pp. 15–23, 2017. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pfi.21727.

[14] B. G. Schor, C. Norval, E. Charlesworth, and J. Singh, "Mind the gap: Designers and standards on algorithmic system transparency for users," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2024.

[15] K. Fischer, H. M. Weigelin, and L. Bodenhagen, "Increasing trust in human–robot medical interactions: effects of transparency and adaptability," *Paladyn, Journal of Behavioral Robotics*, vol. 9, no. 1, pp. 95–109, 2018.

[16] L. Aquilino, P. Bisconti, A. Marchetti, *et al.*, "Trust in ai: Transparency, and uncertainty reduction. development of a new theoretical framework," in *CEUR WORKSHOP PROCEEDINGS*, pp. 19–26, CEUR-WS. org, 2024.

[17] M. Cai, Q. Jin, J. Zhou, and X. Luo, "How Transparency Shapes the Quality of Human-Robot Interaction: An Examination of Trust, Perception, and Workload," *International Journal of Social Robotics*, vol. 17, pp. 1335–1362, July 2025.

[18] M. Rönkkö and E. Cho, "An Updated Guideline for Assessing Discriminant Validity," *Organizational Research Methods*, vol. 25, pp. 6–14, Jan. 2022. Publisher: SAGE Publications Inc.

[19] J. Fenn, C.-S. Tan, and S. George, "Development, validation and translation of psychological tests," *BJPsych Advances*, vol. 26, no. 5, pp. 306–315, 2020.

Measuring Transparency in Intelligent Robots

Georgios Angelopoulos[1,*,+], **Dimitri Lacroix**[2,**,+], **Ricarda Wullenkord**[2], **Alessandra Rossi**[1], **Silvia Rossi**[1], **and Friederike Eyssel**[2]

[1]Interdepartmental Center for Advances in Robotic Surgery - ICAROS, University of Naples Federico II, Naples, 80131, Italy
[2]Center for Cognitive Interaction Technology - CITEC, Bielefeld University, Bielefeld, 33619, Germany
[*]georgios.angelopoulos@unina.it
[**]dimitri.lacroix@uni-bielefeld.de
[+]these authors contributed equally to this work

ABSTRACT

As robots become increasingly integrated into our daily lives, the need to make them transparent has never been more critical. Yet, despite its importance in human-robot interaction, a standardized measure of robot transparency has been missing until now. This paper addresses this gap by presenting the first comprehensive scale to measure perceived transparency in robotic systems, available in English, German, and Italian languages. Our approach conceptualizes transparency as a multidimensional construct, encompassing explainability, legibility, predictability, and meta-understanding. The proposed scale was a product of a rigorous three-stage process involving 1,223 participants. Firstly, we generated the items, secondly, we conducted an exploratory factor analysis, and thirdly, a confirmatory factor analysis served to validate the factor structure of the newly developed TOROS scale. The final scale encompasses 26 items and comprises three factors: *Illegibility*, *Explainability*, and *Predictability*. TOROS demonstrates high cross-linguistic reliability, inter-factor correlation, model fit, internal consistency, and convergent validity across the three cross-national samples. This empirically validated tool enables the assessment of robot transparency and contributes to the theoretical understanding of this complex construct. By offering a standardized measure, we facilitate consistent and comparable research in human-robot interaction in which TOROS can serve as a benchmark.

Introduction

Humans rely on perception, interpretation, and anticipation to interact effectively with their environment, and the same applies when interacting with robots[1]. When a robot behaves in unexpected or unclear ways, users may need to pause, hesitate, or intervene, which can disrupt smooth collaboration, reduce efficiency, or even compromise safety[2]. Ensuring that robot actions are understandable and predictable is therefore crucial for effective interactions. This characteristic is often discussed under the term transparency, which has become a central concern in robotics and artificial intelligence[3].

Transparency is explicitly recognized as an important issue in governance frameworks. The Principles of Robotics issued by the UK's Engineering and Physical Sciences Research Council state that "Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead, their machine nature should be transparent"[4]. At the European level, the Artificial Intelligence Act (EU AI Act) further highlights transparency as a legal requirement. Article 13 stipulates that "AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable providers and users to reasonably understand the system's functioning"[5]. These frameworks demonstrate that transparency is not simply a desirable design principle but a mandated condition for the deployment of AI and robotic systems in society[6,7].

Despite this recognition, a clear and commonly accepted definition of transparency in human-robot interaction (HRI) is still missing. Some authors define transparency in the sense of predictability, such as "Transparency is essentially the opposite of unpredictability" [8, p. 193]. Accordingly, if a robot provides cues that support its users in anticipating its next actions and states, the robot should be deemed more predictable and, thus, transparent. In contrast, other definitions of transparency emphasize the aspect of legibility. Wortham et al. [9, p. 274] defined transparency as "... the extent to which the robot's ability, intent, and situational constraints are understood by users". According to Kim and Hinds [10, p. 81], the notion of explainability is also intertwined with transparency. They propose "Transparency is the robot offering explanations of its actions". These explanations can be provided before an action is performed, during an action, or after performing an action[11,12]. Recent works, however, combine the aforementioned aspects, suggesting that transparency in HRI should encompass what a robot is doing (legibility), why it is doing that (explainability), and what it will do next (predictability)[13–15]. Therefore, we take a multifaceted initial approach to transparency, integrating elements of explainability, legibility, and predictability.

Transparency of robot behavior is essential for several reasons: For end-users, it regulates expectations, fosters trust, acceptance, and improves user experience[16–19]. For developers, it aids processes associated with design, debugging, and evaluation by making robot behavior easier to inspect and adjust[3,20]. However, transparency is a double-edged sword: While transparency can enhance trust and acceptance, it may also result in negative consequences, such as overtrust[21] or information overload. For instance, when explanations provided by the robot are too extensive[22]. Transparency is distinct from related constructs like trust or acceptance, in that it concerns whether the system's functioning can be understood, not whether it is deemed reliable or benevolent (trust), nor enjoyable or useful (acceptance). As such, transparency deserves to be measured as a construct on its own.

Neverthless, measuring transparency remains problematic. Schött et al.[23] have highlighted the absence of consensus regarding the reliability and validity of metrics employed to assess the transparency of a robot's behavior. In addition, Claure et al.[24] and Bartneck et al.[25] underscored the critical need for developing a standardized measurement framework to assess transparency. When measuring transparency, straightforward questions such as "Is the robot transparent?" or "Is the robot predictable?" might seem adequate at first glance[26,27]. However, such face-valid, single-item measures fail to comprehensively capture all facets of a complex construct like transparency, leading to inconsistent and unreliable data[28]. The lack of consensus regarding the measurement of transparency in HRI research suggests that it may be understood differently from one individual to another. Similarly, lay people may perceive transparency differently from HRI researchers and practitioners, who distinguish explainability, legibility, and predictability as dimensions of transparency. For instance, legibility and predictability are not easily distinguishable in terms of human understanding. In fact, human cognition is deeply based on predictive processes to understand the environment and prepare individuals for actions[29,30]. It is then possible that people, when understanding something, have the feeling that it is predictable as well[31]. The development of a validated scale is therefore essential for methodological rigor and for the theoretical understanding of transparency.

The present paper addresses exactly this gap. To the best of our knowledge, it proposes the first measurement instrument to assess the perceived transparency of a robot's behavior. We present the validation process, which follows the state-of-the-art for scale creation[28,32]. This process comprises three distinct stages, as shown in Figure 1, each integral to ensure the robustness and validity of the resulting final scale. These stages consist of (1) Initial item generation to formulate a set of items, (2) Exploratory Factor Analysis (EFA) to discern a factor arrangement, and (3) a Confirmatory Factor Analysis (CFA) to validate the scale's factor structure. Factor analyzes are a crucial step to explore and validate a measurement, as they narrow down a large number of variables (e.g., the items of a scale) to a smaller number of factors (e.g., the expected dimensions of a scale), using patterns of correlation between variables. Whereas EFAs are used to extract factors by grouping correlated variables, CFAs served to test how an already assumed factor structure fit a set of observed variables[33]. The research protocol was approved by Bielefeld University's Ethics Review Board (application number 2023-349). Participants provided informed consent and the research protocol was in accordance with the Declaration of Helsinki.

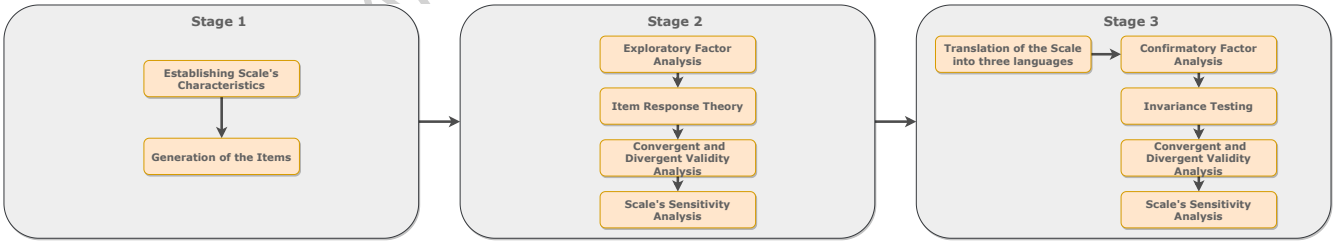


Figure 1. The three stages of the scale development.

Stage 1

Stage 1 of the scale development involved an examination of existing definitions and measures to allow the creation of the first version of a scale assessing a robot's perceived transparency. A literature review suggested that a robot's behavior requires three characteristics to be transparent: explainability, legibility, and predictability[13–15]. However, transparency also emerges from how these dimensions help the human perceiver to build a mental model of the robot's functioning[34,35]. We refer to this fourth aspect as meta-understanding. With these four characteristics established, we generated items that effectively capture each aspect of transparency. Here, we were guided by the methodological recommendation to generate three to five times the number of items intended for a resulting measurement instrument, with at least four items per dimension[32]. Consequently, 64 items in total were initially developed and distributed evenly across the four dimensions. Moreover, since no validated transparency scale is available in the literature yet, the first version of the scale was designed so items from previous studies that

explicitly measured specific aspects of transparency (e.g., *"I find the robot's behavior easy to understand"*) were incorporated, as explained in recent HRI and Human-Computer Interaction (HCI) literature[15,23,36]. Each item was formulated to carefully match the theoretical notions associated with explainability, legibility, predictability, and meta-understanding. This way, we made sure that the items were relevant and sufficient to fully capture each dimension.

Furthermore, in Stage 1, we deliberately created multiple items with similar content or phrasing. The rationale behind this strategy was two-fold. First, it allowed for a more comprehensive coverage of the construct's content domain, ensuring that various nuances and aspects of each dimension are captured. Second, it provides the opportunity to empirically evaluate which items perform best in terms of psychometric properties during subsequent phases of scale refinement. By including redundant items at this stage (e.g., *"The robot's behavior is predictable"*, *"The robot's actions are unpredictable"*), we increase the likelihood of identifying the most robust and discriminating items for the final scale, potentially improving its overall reliability and validity[37].

In addition, to capture the different ways in which individuals might perceive and describe their experiences with robots, a mix of personal and nonpersonal statements was developed. Personal statements (e.g., *"I can [...]", "I find [...]", "I feel [...]"*) refer to direct expressions of the participant's experiences and feelings. Nonpersonal statements (e.g., *"The robot's behavior", "The robot's explanation [...]", "The robot's actions"*) provide an assessment of the robot's characteristics. For each expected dimension of transparency, we complemented these regular items (i.e., statements phrased in the direction of the targeted constructs, also known as positively worded items)[38,39] with three reverse-worded items (asking for agreement with statements phrased in the opposite direction of the targeted construct)[39–41], each for personal and non-personal statements (e.g., *"I find it difficult [...]", "The robot does not [...]"*). This was done to counter-response the acquiescence bias (i.e., tendency of respondents to use agreement with statement as the default answer to spare effort), careless responding, and to increase the reliability of responses[40]. Responses are measured using a 7-point Likert scale, ranging from *"Strongly Disagree" (1)* to *"Strongly Agree" (7)*. Each answer point is fully labelled, which increases the understandability of the items, reduces cognitive load of respondents, and limitates the risk of misresponse to reverse-worded items[38]. 7-point Likert scales offer an optimal balance between ease of use, adjustment to memory span, and accuracy[42]. In addition, as 7-point Likert scales include a midpoint, they allow to identify answers that are neutral (i.e., lacking strong positive or negative evaluations) or ambivalent (i.e., evaluations that are both strongly positive and negative), which can be valuable leads for further research[43–45]. Even-numbered response formats (e.g., 4- or 6-point scales) were not considered because they prevent participants from expressing neutrality or ambivalence by constraining them to make a choice (i.e., "agree" or "disagree"). Doing so, even-numbered response formats can lead to answers that do not reflect actual evaluations of respondents[38,42,46], and can elicit negative reactions from ambivalent participants[38,47]. However, prior research has advocated for the use of even-numbered response formats because midpoints in odd-numbered response formats are deemed a threat to reliability[42,46]. Therefore, and despite inconsistent evidence for such an effect[42,46], it was taken into account when assessing the reliability of the scale at the next Stages. The initial 64 items can be found in Table S1 of the Supplementary Information.

Stage 2

Following item generation, Stage 2 of the scale development and validation process featured an empirical experimental study. This enabled an exploratory factor analysis to examine the underlying structure of the scale. Such analysis served to confirm the hypothesized four-factor model of perceived transparency. EFA was used because it is instrumental in assessing the scale's factorial structure and facilitates item reduction by identifying items that strongly contribute to each factor. At this stage, an assessment of the scale's item difficulty and discrimination parameters was also included. Additionally, Stage 2 served to evaluate the scale's convergent and divergent validity through controlled manipulations of the primary dimensions of transparency (explainability, legibility, and predictability), as explained in previous works[13–15], with image vignettes (controlled visual scenarios). This process was crucial in order to confirm whether changes in these dimensions correspond to variations in perceived transparency. That way, we can provide evidence for the scale's sensitivity and construct validity.

Before Experiment 1, we conducted a pretest to identify a hypothetical everyday life scenario featuring HRI that would effectively discriminate transparency. Thereby, we maximized the efficiency of our experimental manipulation in testing the scale's sensitivity and effectiveness for measuring transparency. Based on the pretest's findings, we selected a scenario where a robot heading towards a charging station to refill its battery. The results of the pretest can be found in the Supplementary Information.

Experiment 1 employed a $2 \times 2 \times 2$ between-subject design which manipulated the explainability, legibility, and predictability of a robot's behavior as either low or high, using eight image vignettes, selected based on the results of the pretest. Experiment 1 was implemented on Qualtrics. Participants were presented with the purpose and procedure of the experiment. Those who gave their informed consent were randomly assigned to one of the eight between-subject conditions resulting from the manipulation of three variables related to transparency. After viewing the scenario, participants were required to self-report the following

dependent variables: Perceived transparency with 64 items developed during Stage 1 (Item Generation), trust towards robots with 20 items from the Multi-Dimensional Measure of Trust (MDMT)[48], and acceptance of robots using seven subscales from a toolkit based on the Unified Theory of Acceptance and Use of Technology (UTAUT), suggested for HRI research[49]. The order of the items within each scale was randomized. This was done to avoid a potential order effect induced by item presentation[50], and to vary the alternance of regular and reverse-worded items[40]. Additionally, demographic questions (i.e., age, gender, education, self-assessed English language proficiency) and prior experience with robots were assessed using a scale based on[51]. Finally, two attention checks were included, one at the beginning and one at the end of the study. Only complete datasets from participants over 18 years of age and with a self-declared English proficiency at the A2 level (Elementary) and above were included. Data from people failing both attention checks were excluded. The data collection was planned to conclude after obtaining complete datasets from 320 participants. This sample size was strategically chosen to meet the requirement of having at least 5 participants per item for factor analysis[52], considering the transparency scale consisted of 64 items. Additionally, the sample size met the recommended threshold of 300 participants for robust factor analysis[33]. Participants for Experiment 1 were recruited via Prolific, and they were reimbursed with £1.50 for participating. The pre-registration for Experiment 1 is available at https://aspredicted.org/ZTT_STV.

Sample

For Experiment 1, we recruited 371 participants. Following the preregistered exclusion criteria, 49 participants were removed from the sample; 13 did not complete the study, and 36 were excluded because they did not pass both attention checks. Thus, the final sample comprised $N = 322$ participants, slightly exceeding our pre-registered target by two participants. The demographic breakdown, as detailed in Table S3 of the Supplementary Information, was as follows: 182 females, 135 males, two identifying as diverse, two who preferred not to disclose their gender, and one gender-fluid participant. The participants' age range was between 18 and 71 years ($M = 29.820$, $SD = 9.340$).

Results

An EFA using *SPSS 28.0* was used to assess the factorial validity of the 64-item scale. This analysis utilized Principal Axis Factoring (PAF) with an oblique (direct-oblimin) rotation, which allows intercorrelations among factors[53,54]. This approach was accompanied by evaluations of univariate statistics, the Kaiser-Meyer Olkin (KMO) measure of sampling adequacy, Bartlett's test of sphericity, eigenvalues, and the Scree plot.

The KMO for the sample data was .976. Consequently, the data appeared suitable for an EFA[55,56]. Bartlett's test further validated the suitability of this statistical method, producing a $\chi^2 = 18084.169$, df $= 2016$, $p < .001$. The Scree plot, in Figure 2, revealed a clear breakpoint after two factors, suggesting one major factor. However, the curve flattened after five factors, suggesting that the optimal number of factor to retain was four to five[33]. Eigenvalues above 1.0 suggested seven factors. At this stage of the scale development, we opted for a five-factor model (instead of one to four) to avoid discarding potentially meaningful factors before further examination of the items. In addition, the five-factor model was preferred over a seven-factor model to avoid a risk of model overfit. The five factors we have retained accounted for 64.154% of the variance. This result demonstrated the strong explanatory power of the scale.

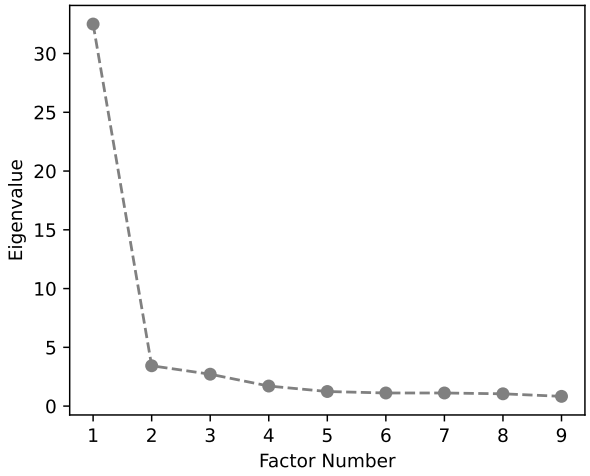


Figure 2. Scree plot illustrating the eigenvalues of extracted factors of the scale in Experiment 1.

Initially, all 64 items were retained for the initial EFA. The results, detailing the factor loadings, are depicted in Tables S5 and S6 of the Supplementary Information. Subsequent iterations of the EFA were conducted using the same settings, with item reduction guided by specific criteria to ensure a robust factor structure. These criteria involved removing items with loadings below .400 to maintain a strong factor representation and enhance interpretability[57] . Additionally, items loaded on multiple factors with a loading difference of less than .100 were excluded to minimize ambiguity in item association[57]. Furthermore, factors containing fewer than three items were removed to ensure reliable and meaningful factor solutions[28,58].

Initial item removals included five rounds of EFA, in which low factor loadings, multiple high factor loadings, and factor loadings on underrepresented factors were taken into account as exclusion criteria. The final EFA resulted in a refined 31-item scale structured into three distinct factors. The revised factor model demonstrates robust factor loadings. The detailed factor loadings for this definitive structure are presented in Tables S7 and S8 of the Supplementary Information.

Following the iterations of EFA, an inter-item correlation analysis was conducted to minimize redundancy within each factor. Items with excessively high correlations (greater than .700 with more than one item) were removed to ensure the distinctiveness of each item within the factors. This approach confirmed that the remaining items clearly and distinctly represented the underlying constructs without undue overlap. At first, the inter-item correlations for Factor 1 were analyzed, with correlation coefficients ranging from .511 to .793 with a mean of $M = .680$. Table S9 of the Supplementary Information shows that four items had correlation coefficients higher than .700 with more than one variable, leading to their removal. Table S10 of the Supplementary Information shows the final inter-item correlation matrix of Factor 1 ranging from .511 to .728 with a mean of $M = .642$. Afterward, we analyzed Factor 2 (as shown in Table S11 of the Supplementary Information), in which the correlation coefficients ranged from .479 to .740 with $M = .607$. One item had correlation coefficients greater than .700 with more than one item and was removed. A new analysis was conducted (depicted in Table S12 of the Supplementary Information), showing that the correlation coefficients ranged from .479 to .690 with a mean of $M = .598$. Finally, as depicted in Table S13 of the Supplementary Information, the values of the inter-item correlation for Factor 3 ranged from .490 to .683 with a mean of $M = .572$. No items were removed, leading to a 26-item scale.

To validate the robustness and appropriateness of the factor structure of the 26-item scale, we checked sampling adequacy, factorability, internal consistency, and convergent validity. The KMO reaffirmed the suitability of the data for EFA with a value of .968. Bartlett's test supported the factorability of the correlation matrix, yielding a $\chi^2 = 6217$, df $= 325$, $p < .001$. The new factor loadings for both pattern and structure matrices were examined to confirm the appropriateness of the items within each factor. Internal consistency was assessed using Cronbach's Alpha (α) and McDonald's Omega (ω). All factors demonstrated high internal consistency, with Factor 1 recording values of .930 for both Cronbach's Alpha and McDonald's Omega, Factor 2 showing a Cronbach's Alpha of .926 and McDonald's Omega of .927, and Factor 3 similar to Factor 1 with values of .930 for both metrics. The high internal consistency of each factor suggested that the scale is reliable, especially when accounting for a potential negative impact of a 7-point format on it[46]. Composite Reliability (CR) and Average Variance Extracted (AVE) were calculated to evaluate the factors' convergent validity. The CR values for Factor 1, Factor 2, and Factor 3 were satisfactory at .800, .800, .735, respectively. Regarding the AVE, Factor 1, Factor 2, and Factor 3 recorded values of .438, .438, and .463, respectively. Despite the AVE values being marginally below the ideal threshold of .5, the high CR values justify proceeding with the utilization of the scale[59]. The comprehensive results from these evaluations are presented in Table 1.

Moreover, to further evaluate the 26-item scale, we employed Item Response Theory (IRT). We did so to analyze item difficulty and dicrimination[60], as outlined in Table S14 of the Supplementary Information. IRT is a statistical approach that models the relationship between an individual's response to an item and their level of the underlying construct being measured, known as the latent trait[32,60,61]. To do so, IRT conceptually distinguishes a respondents' abilities or attitudes that determine their response to an item from the characteristics of the item, assuming both determine an observed score[60,61]. The IRT model results are presented in Figure 3 using an Item Characteristic Curve (ICC) for all 26 items. Such ICC featured a sigmoid shape, indicating a good fit between the model and the empirical data[60]. Specifically, the sigmoid pattern suggests that as the latent trait increases, the probability of a correct response also increases systematically, reflecting an effective measurement scale. For item difficulty, values ranged from .580 to .710 ($M = .660$, $SD = .040$), which is within the acceptable range (i.e., .200 - .900) and slightly above the optimal range (i.e., .500 - .600, with some authors arguing for .500 - .750)[62–64]. As higher values imply lower item difficulty, the scale can be deemed moderately difficult[62–64]. Regarding item discrimination, the range was between .660 and 0.810 ($M = .745$, $SD = .045$), with values above .350 to .600 representing excellent levels of item discrimination[64,65]. Both item difficulty and item discrimination confirm that the items are effective in differentiating between individuals. These results discard a common concern in psychometrics regarding the potentially negative effect of reverse-worded items on the reliability and understandability of a scale[39,66]. Indeed, neither Factor 1 (8 reverse-worded items out of 10) nor Factor 3 (2 reverse-worded items out of 9) are noticeably worse (or better) in terms of reliability, difficulty or discrimination than Factor 2 (no reverse-worded items). It must be noted that the higher number of reverse-worded items in Factor 1 compared to the other factors can be due to an artefact called a method factor. Indeed, regular and reverse-worded items tend to load on different

Item	Factor Pattern Loadings			Factor Structure Loadings	Internal Reliability		Convergent Validity	
	Factor 1	Factor 2	Factor 3		α	ω	CR	AVE
The robot's overall functioning is a mystery to me.	**-.797**	-.008	.032	-.683				
It is hard to make sense of the robot's general functioning.	**-.770**	-.032	.058	-.656				
It is difficult to get a clear picture of the robot's overall operations.	**-.706**	-.101	.021	-.690				
I am confused about the robot's general objectives.	**-.692**	-.144	.020	-.715				
I am unsure what the robot does.	**-.688**	-.098	-.038	-.724	.930	.930	.800	.438
I cannot comprehend the robot's inner processes.	**-.673**	.004	-.056	-.641				
I cannot explain the robot's behavior.	**-.671**	.135	-.284	-.729				
It is impossible to know what the robot does.	**-.638**	-.049	-.124	-.715				
It is clear to me what the robot does.	**.467**	.347	.106	.799				
I have a clear understanding of how the robot operates in general.	**.426**	.320	.067	.739				
I feel like the robot's explanations are useful.	.024	**.832**	-.039	.688				
The robot explains complex tasks in a way that is easy to understand.	.119	**.753**	-.024	.718				
The robot provides detailed explanations of its actions.	.008	**.708**	.084	.678				
The robot provides clear explanations for its actions.	.067	**.657**	.163	.757	.926	.927	.865	.481
The robot's explanations for its actions are straightforward.	.049	**.654**	.223	.792				
I feel informed about the robot's activities.	.251	**.625**	.044	.787				
The robot conveys its overall state effectively.	.096	**.600**	.110	.688				
It is easy for me to foresee the robot's future actions.	-.043	-.011	**.829**	.685				
The robot's behavior is predictable.	-.034	.090	**.785**	.739				
I feel confident in predicting the robot's next moves.	.057	.064	**.732**	.751				
It is easy to anticipate what will follow the robot's behavior.	.057	.038	**.723**	.721				
It is difficult for me to tell what the robot will do next.	-.318	.168	**-.666**	-.728	.930	.930	.735	.463
The robot's next steps are clear to me.	.049	.209	**.656**	.798				
The robot's actions are obvious.	-.018	.266	**.585**	.725				
The robot provides cues that help predict its next actions.	-.082	.336	**.554**	.700				
The robot's behavior does not help predict what it will do next.	-.277	.076	**-.536**	-.654				

Table 1. The scale after the removal of items with low loadings and high inter-item correlation in Experiment 1.

factors in factor analyses, even when they are supposed to measure the same construct[39,41,66]. The presence of such an artefact in the three-factor structure of the scale is examined at the next stage.

At the end of the item removal procedure and scale evaluation, the 26-item scale was composed of three factors. Factor 1 encompassed ten items (two regular, eight reverse-worded) that were assumed to capture perceived legibility (e.g.,"It is impossible to know what the robot does") and meta-understanding (e.g., "I cannot comprehend the robot's inner processes"). Overall Factor 1 pertains to a perceived difficulty to understand a robot's behavior, hindering the possibility to infer its functioning, goals, or underlying processes. Factor 2 comprised seven regular items that were supposed to capture perceived explainability of the robot (e.g., "The robot provides detailed explanations of its actions"). A notable exception is "The robot conveys its overall state effectively", that was supposed to measure meta-understanding. However, this item probably loaded on items pertaining to explainability because "conveys" implied the ability of the robot to communicate about a "state" that guided its behavior. Finally, Factor 3 included nine items (seven regular, two reverse-worded) that were assumed to capture perceived predictability, with one supposed to measure perceived legibility (i.e., "The robot's actions are obvious"). Regarding the latter, it is possible that a robot's actions were deemed "obvious" when they seemed to foreshadow the subsequent ones. These results are consistent with the three-factor approach of transparency: Legibility (Factor 1), Exainability (Factor 2), Predictability (Factor 3) [13–15]. However, Factor 1 mostly included reverse-worded items, and thus tapped more into the polar opposite of legibility: Illegibility. In addition, illegibility hinted more at how a robot's actions fail to infer internal states, such as goals or functioning, rather than how the actions themselves are understandable. Hence the presence of items hypothetically related to both legibility and meta-understanding. We decided to keep the factors unlabelled until Experiment 2 confirmed these results.

Prior to assessing the sensitivity of the 26-item scale to an experimental manipulation of transparency of a robot's behavior, we examined the reliability of the other questionnaires (i.e., MDMT, UTAUT, and experience with robots). Cronbach's alpha ($> .7$) confirmed their internal consistency, as demonstrated in Table S15 of the Supplementary Information. Afterward, we

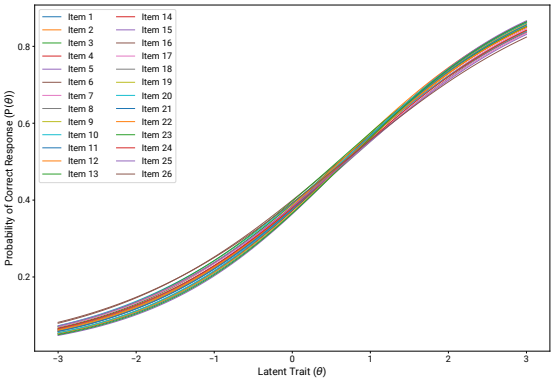


Figure 3. The Item Characteristic Curve exhibits a sigmoid shape.

conducted a series of 2x2x2 factorial ANCOVAs. A first set of ANCOVAs was conducted to verify that our measurement of perceived transparency of robots' behavior is sensitive to a manipulation of the dimensions of transparency identified by the available literature. Therefore, manipulated levels of explainability, legibility, and predictability (each set at low versus high) as the independent variables while utilizing the means for each factor as well as the mean of all items from the 26-item scale as the dependent variables. A second set of ANCOVAs was dedicated to confirm that transparency of a robot positively influences acceptance and trust towards it. Thus, levels of explainability, legibility, and predictability were used as independent variables, with performance and moral trust (MDMT), acceptance, and anxiety (UTAUT) as dependent variables. In cases where experience with robots was significantly correlated with the tested dependent variable (see Figure 4), it was included as a covariate in the ANCOVA. Table 2 shows the results of all the ANCOVAs.

We found a significant main effect of manipulated explainability, legibility, and predictability on transparency calculated using the average of all the items. However, no interaction effect between manipulated explainability, legibility, and predictability was identified. Similarly, we obtained significant main effects of manipulated explainability, legibility, and predictability on Factor 1, Factor 2, and Factor 3. However, no interaction effect between manipulated explainability, legibility, and predictability on Factor 1, Factor 2, or Factor 3 was observed. Prior experience with robots was a significant covariate for average perceived transparency, Factor 1, Factor 2, and Factor 3. These results confirmed the sensitivity of our measurement of perceived transparency to an experimental manipulation of the transparency of a robot's behavior.

Regarding trust towards the robot, we found significant main effects of manipulated explainability and legibility on performance (see Table S16 of the Supplementary Information) and moral trust towards the robot (see Table S17 of the Supplementary Information). However, although the main effect of manipulated predictability on performance trust was significant, we did not find a significant effect of manipulated predictability on moral trust towards the robot. No significant interaction effect between manipulated explainability, legibility, and predictability on performance trust or moral trust was discovered. Prior experience with robots was a significant covariate for performance trust and moral trust towards the robot. Finally, we observed a significant main effect of manipulated explainability, legibility, and predictability on acceptance of the robot (see Table S18 of the Supplementary Information). Prior experience with robots was a significant covariate for acceptance. No interaction effect between manipulated explainability, legibility, and predictability on acceptance was identified. No significant effect of the independent variables on anxiety was found (see Table S19 of the Supplementary Information). These results provided support for the positive effect of trnasparency of a robot's behavior on acceptance and trust towards robots.

Stage 3

Following Stage 2, which concluded with a 26-item scale, we proceeded to Stage 3 of the scale development process. In this stage, we designed Experiment 2 to validate the factor structure. The first step involved translating the 26-item scale into German and Italian, employing a forward and backward translation process by two independent translators[67]. This was essential for testing the scale's properties across languages and for ensuring cross-linguistic reliability and validity. Additionally, the effectiveness of the scale was assessed by measuring participant responses under conditions of low versus high transparency with video vignettes. These vignettes were based on the scenarios of Stage 2, and the video format was meant to have the transparency of the robot assessed in more ecological settings (i.e., after observing a real robot in action). Additionally, Stage 2

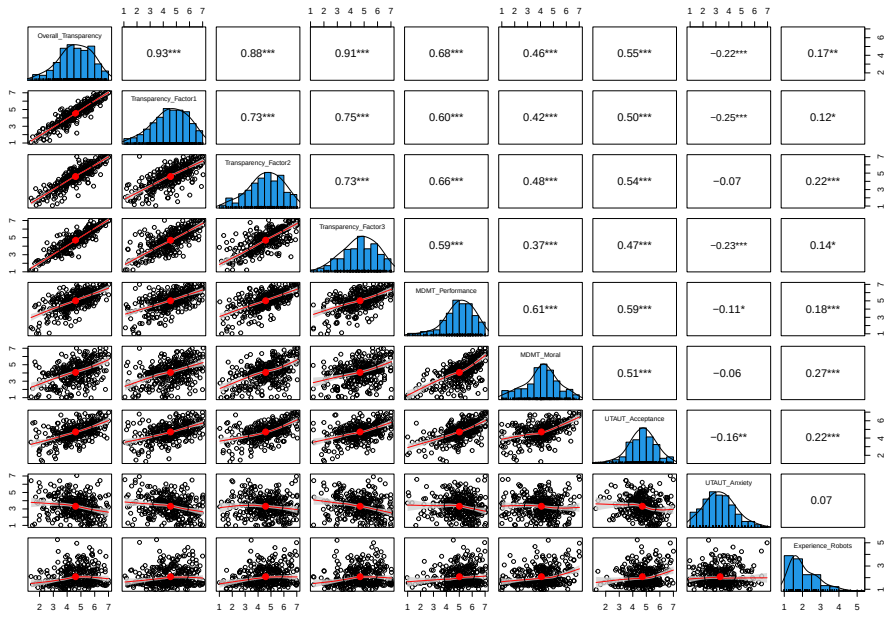


Figure 4. Correlation matrices between the dependent and control variables of Experiment 1.

included a CFA to validate the factor structure of the 26-item scale identified in the previous Stage. As a result, Stage 3 was critical in demonstrating the scale's capacity to discriminate between different levels of perceived transparency, demonstrating its practical applicability and psychometric consistency across diverse cultural contexts, as in previous studies[68].

Experiment 2 was designed as a 2 (low and high transparency conditions) ×3 (3 languages) between-subject experiment where that served to manipulate the transparency of a robot's behavior via video vignettes featuring the robot Pepper (Softbank Robotics). Given that explainability, legibility, and predictability all had a significant influence on each factor of the scale in the previous stage, for Stage 2 we used two conditions to manipulate the overall transparency. Moreover, Experiment 2 was conducted in three languages (English, German, and Italian). Participants were presented with the purpose and procedure of the experiment. Those who gave informed consent were randomly assigned to one of the two transparency conditions, and the corresponding video was displayed. To ensure that participants watched the entire video featuring the robot behaving high vs. low in transparency at least once, the "Continue" button appeared only after approximately 5 seconds. Following the video, participants completed the resulting refined scale from Stage 2, followed by the MDMT scale[48] and the subscales from the UTAUT toolkit[49]. Translated versions of these scales, prepared by language experts, were used for German and Italian participants. The order of the items for each scale was randomized. Additionally, demographic questions (i.e., participants' age, gender, education, self-assessed English language proficiency) and prior experience with robots were assessed using a scale based on[51]. Finally, two attention checks were included at the study's beginning and end, and one memory check was included after the video's projection. Only complete datasets from participants over 18 years of age and with a self-declared language level (English, German, or Italian, depending on the language condition) above B2 level (Upper Intermediate) were included. Data from people failing the attention checks or the memory check were excluded. Following existing recommendations[33], the data collection was planned to conclude after obtaining complete datasets from 300 participants per sub-sample, resulting in a total number of 900 complete data sets. The Study 2 pre-registration details are available at https://aspredicted.org/JMC_3B8.

Sample

We recruited 927 participants using Prolific. Following our pre-registered exclusion criteria, 26 participants were disqualified; 17 failed to complete the study, 1 was excluded for not passing the video attention check, and 8 were removed for insufficient language proficiency. The final sample comprised $N = 901$ participants. As detailed in Table S20 of the Supplementary Information, the demographics were as follows: 427 females, 452 males, 14 identifying as diverse, and 8 who preferred not to disclose their gender. The age range of participants was between 18 and 71 years ($M = 37.607$, $SD = 12.215$), and their experience with robots was rated below average ($M = 1.930$, $SD = 1.160$). The average duration of participation was recorded at 9.1 minutes, and each participant were compensated with £1.50 for their participation in the study.

Parameter	df	SS	MS	F value	p value	ηp^2	95% CI
Dependent variable: Overall Score of the Transparency							
Explainability condition	1	74.10	74.13	70.53	<.001***	.18	[0.12, 1.00]
Legibility condition	1	9.20	9.23	8.78	.003**	.03	[0.01, 1.00]
Predictability condition	1	27.70	27.67	26.33	<.001***	.08	[0.04, 1.00]
Prior experience with robots	1	17.60	17.64	16.78	<.001***	.05	[0.02, 1.00]
Explainability: Legibility	1	0.70	0.70	0.67	.415	<.01	[0.00, 1.00]
Explainability: Predictability	1	0.30	0.29	0.28	.599	<.01	[0.00, 1.00]
Legibility: Predictability	1	1.90	1.90	1.803	.180	.01	[0.00, 1.00]
Explainability: Legibility: Predictability	2	0.60	0.62	0.59	.444	<.01	[0.00, 1.00]
Residuals	313	329.0	1.05				
Dependent variable: Factor 1							
Explainability condition	1	67.50	67.50	46.55	<.001***	.13	[0.08, 1.00]
Legibility condition	1	13.30	13.35	9.21	.003**	.03	[0.01, 1.00]
Predictability condition	1	31.80	31.85	21.96	<.001***	.07	[0.03, 1.00]
Prior experience with robots	1	12.50	12.48	8.60	.004**	.03	[0.01, 1.00]
Explainability: Legibility	1	0.90	0.94	0.65	.421	<.01	[0.00, 1.00]
Explainability: Predictability	1	0.10	0.07	0.05	.823	<.01	[0.00, 1.00]
Legibility: Predictability	1	2.80	2.83	1.95	.163	.01	[0.00, 1.00]
Explainability: Legibility: Predictability	2	1.60	1.58	1.09	.298	<.01	[0.00, 1.00]
Residuals	313	453.90	1.45				
Dependent variable: Factor 2							
Explainability condition	1	116.50	116.45	91.57	<.001***	.23	[0.16, 1.00]
Legibility condition	1	6.60	6.57	5.17	.024*	.02	[0.00, 1.00]
Predictability condition	1	17.30	17.32	13.62	<.001***	.04	[0.01, 1.00]
Prior experience with robots	1	32.60	32.58	25.61	<.001***	.08	[0.04, 1.00]
Explainability: Legibility	1	1.70	1.73	1.36	.244	<.01	[0.00, 1.00]
Explainability: Predictability	1	0.10	0.12	0.10	.756	<.01	[0.00, 1.00]
Legibility: Predictability	1	1.00	1.00	0.78	.377	<.01	[0.00, 1.00]
Explainability: Legibility: Predictability	2	0.10	0.06	0.05	.824	<.01	[0.00, 1.00]
Residuals	313	398.10	1.27				
Dependent variable: Factor 3							
Explainability condition	1	54.00	54.04	41.52	<.001***	.12	[0.07, 1.00]
Legibility condition	1	7.40	7.42	5.70	.018*	.02	[0.00, 1.00]
Predictability condition	1	32.40	32.37	24.87	<.001***	.07	[0.03, 1.00]
Prior experience with robots	1	14.20	14.20	10.91	.001**	.03	[0.01, 1.00]
Explainability: Legibility	1	0.10	0.10	0.08	.781	<.01	[0.00, 1.00]
Explainability: Predictability	1	2.50	2.52	1.94	.165	.01	[0.00, 1.00]
Legibility: Predictability	1	1.80	1.77	1.36	.244	<.01	[0.00, 1.00]
Explainability: Legibility: Predictability	2	0.50	0.46	0.36	.551	<.01	[0.00, 1.00]
Residuals	313	407.40	1.30				

Note: df = Degrees of freedom; SS = Sum of Squares; MS = Mean Squares;
*p< .05; **p< .01; ***p< 0.001;

Table 2. Results of the 3-way ANOVA with the manipulation of explainability, legibility, and predictability of the robot's behavior as independent variables in Experiment 1

Results

A Confirmatory Factor Analysis using the *lavaan* package in R was employed to verify the factor structure of the 26-item scale. The CFA results for the English, German, and Italian versions of the scale indicated a good model fit for the English and Italian versions, with Comparative Fit Index (CFI) values of 0.930 and 0.943, respectively, and Tucker-Lewis index (TLI) values of 0.924 and 0.937. The Root Mean Square Error of Approximation (RMSEA) values were 0.073 for English and 0.067 for Italian, both within acceptable ranges. The Standardized Root Mean Square Residual (SRMR) values were 0.053 and 0.045, respectively, indicating a good fit. The German version showed weaker fit indices with a CFI of 0.890, TLI of 0.879,

RMSEA of 0.093, and SRMR of 0.062, suggesting a less optimal fit than the other versions. Despite this, all factor loadings were significant across all three language versions, indicating that the items loaded well onto their respective factors. Table 3 shows the specific analysis results.

Three alternative Confirmatory Factor analyses were conducted. First, we ran a two-factor CFA where one factor contained all regular items whereas the other factor contained all the reverse worded items. It was conducted to assess the possibility that reverse-worded items in the scale did not create a method factor. Indeed, reverse-worded items can create an artefact in a factor extraction process, which results in the separation of regular and reverse-worded items supposed to measure a single construct in two factors. The factor containing reverse-worded items is referred to as a method factor [39]. The second alternative CFA assumed a single factor to determine whether the scale can be used to calculate a valid transparency score in lieu of the three subscores, each characterizing a dimension of transparency. For both the two-factor and one-factor CFA, most fit indices did not reach acceptable thresholds, with some of them being marginally acceptable. The third alternative CFA was based on our initial assumption according to which transparency encompasses legibility, explainability, predictability, and meta-understanding. The fit indices from this four-factor model ranged from marginally acceptable to excellent. However, apart from the Goodness of Fit Index of the four-factor model in German language, all fit indices indicated a slightly lower fit than the three-factor model. The comparative fit indices of these CFA for every version of the scale can be consulted on Tables S21-23. Taken together, the CFAs revealed that the three-factor solution had the best fit indices in each tested language and that Factor 1 was likely not a method factor.

Standards	English	German	Italian	Acceptable	Excellent
Minimum fit function chi-square (χ^2)	774.253	1067.055	694.216	-	-
Degrees of freedom (df)	296	296	296		
χ^2/df	2.620	3.600	2.350	< 5.00	< 3.00
GFI	0.828	0.756	0.841	> 0.80	> 0.90
RMSEA	0.073	0.093	0.067	< 0.08	< 0.06
AGFI	0.797	0.711	0.811	> 0.80	> 0.90
NFI	0.892	0.854	0.905	> 0.85	> 0.90
CFI	0.930	0.890	0.943	> 0.90	> 0.95
TLI	0.924	0.879	0.937	> 0.90	> 0.95
IFI	0.930	0.900	0.940	> 0.90	> 0.95
SRMR	0.053	0.062	0.045	< 0.08	< 0.05

Table 3. Comparative Fit Indices for English, German, and Italian Models of Experiment 2.

Following the individual CFAs, measurement invariance testing was done to assess the comparability of the scale across the three language versions. Results indicated configural and metric invariance, but not scalar or residual invariance. This suggests that while the construct and factor loadings are comparable across English, German, and Italian, there are differences in item intercepts and residual variances. Specifically, the configural invariance model, which tests if the basic structure is similar across groups, showed a good fit with a CFI of 0.910, TLI of 0.902, and RMSEA of 0.078, indicating that the factor structure is consistent across languages. The metric invariance model, which constrains factor loadings to be equal across groups, also showed acceptable fit (CFI of 0.906, TLI of 0.907, and RMSEA of 0.082), though the chi-squared difference test was significant, $\Delta\chi^2 = 120.47$, $p < 0.001$). The scalar invariance model, which additionally constrains item intercepts to be equal across groups, showed a significant deterioration in fit compared to the metric model, $\Delta\chi^2 = 283.93$, $p < 0.001$, although the overall fit indices remained similar (CFI = 0.906, TLI = 0.907, RMSEA = 0.082). This suggests that while the overall structure and factor loadings are comparable, there may be differences in how participants from different language groups interpret or respond to specific items on the scale. The residual invariance model, which further constrains residual variances to be equal across groups, likewise showed a significant decrease in fit, $\Delta\chi^2 = 226.99$, $p < 0.001$, with a slight drop in CFI to 0.898. This indicates that the unexplained variance in item responses may differ across language groups. Table 4 shows the results of the invariance testing.

After examining scale invariance, it was crucial to assess the scale's convergent validity and internal consistency across the three languages to ensure that the constructs are consistently and accurately measured. Internal consistency was robust across all languages, with Cronbach's Alpha values ranging from .920 to .951 and McDonald's Omega values from .940 to .960, showing high internal consistency for all factors. CR values for all factors exceeded the acceptable threshold of .70, ranging from .760 to .951, ensuring good convergent validity. Additionally, AVE values, which ranged from .547 to .736, consistently exceeded the .50 threshold, showing an improvement from Experiment 1 and confirming that a substantial portion of variance is explained by the factors across all languages. The results are depicted in Table 5.

Model	χ^2	df	$\Delta\chi^2$	Δdf	CFI	TLI	RMSEA	SRMR
Configural	2535.5	888			0.910	0.902	0.078	0.071
Metric	2656.0	934	120.47*	46	0.906	0.907	0.082	0.071
Scalar	2939.9	980	283.93*	46	0.906	0.907	0.082	0.071
Residual	3166.9	1032	226.99*	52	0.898	0.903	0.083	0.070

Note: df = Degrees of freedom; $\Delta\chi^2$ = Change in Chi-Square;
Δdf = Change in Degrees of Freedom
*p< .05; **p< .01; ***p< 0.001;

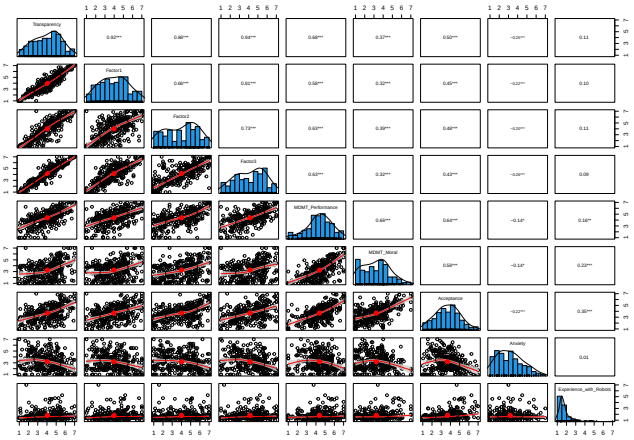
Table 4. Invariance Testing Results for Configural, Metric, Scalar, and Residual Models in Experiment 2

Factor	No. of Items	Internal Consistency						Convergent Validity					
		English		German		Italian		English		German		Italian	
		α	ω	α	ω	α	ω	CR	AVE	CR	AVE	CR	AVE
Factor 1	10	.940	.950	.920	.950	.930	.940	.849	.599	.760	.547	.817	.554
Factor 2	7	.945	.946	.945	.946	.951	.953	.945	.712	.875	.590	.951	.736
Factor 3	9	.950	.960	.950	.960	.950	.960	.849	.669	.860	.668	.878	.686

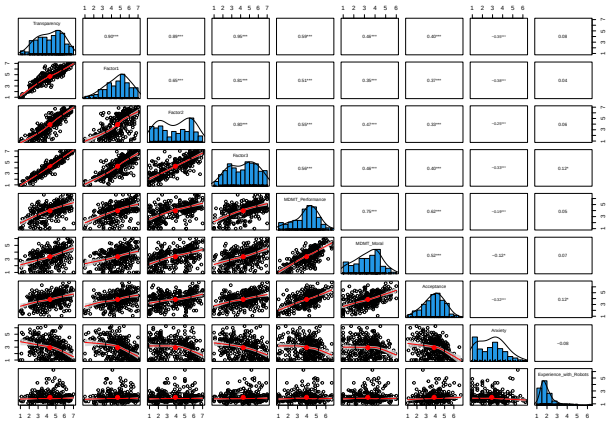
Table 5. Internal consistency and convergent validity for each factor across the scale's English, German, and Italian versions tested in Experiment 2.

Moreover, we examined how the experimental manipulation of a robot's behavior transparency affected participants' perceptions. We used two-tailed t-tests and two-way ANOVAs to analyze the data, considering both the experimental manipulation and the participants' language (English, German, Italian) as factors. We initially planned to include participants' prior experience with robots as a covariate, but found that language significantly influenced this experience ($F(1,895) = 5.32, p = .005, \eta_p^2 = .01, 95\%$ CI $[0.00, 1.00]$), making it unsuitable as a covariate. Results (as shown in Table 6) revealed that participants perceived the robot's behavior as significantly more transparent in the high transparency condition compared to the low transparency condition ($t(874) = 29.88, p < .001, 95\%$ CI $[1.88, 2.14], d = 1.99$). An ANOVA confirmed this main effect of the transparency manipulation. Additionally, we found a significant main effect of language on perceived transparency. When comparing language groups, we found that German participants perceived the robot as more transparent than English participants ($t(598) = 2.55, p = .031, 95\%$ CI $[0.07, 0.53], d = 0.21$) (Post-hoc t-tests with Bonferroni correction). However, we did not obtain significant differences between English and Italian participants ($t(598) = -2.18, p = .086, 95\%$ CI $[-0.48, -0.03], d = -0.18$), or between German and Italian participants ($t(598) = 0.39, p = 1, 95\%$ CI $[-0.18, 0.27], d = 0.03$). There was no significant interaction effect between transparency manipulation and language.

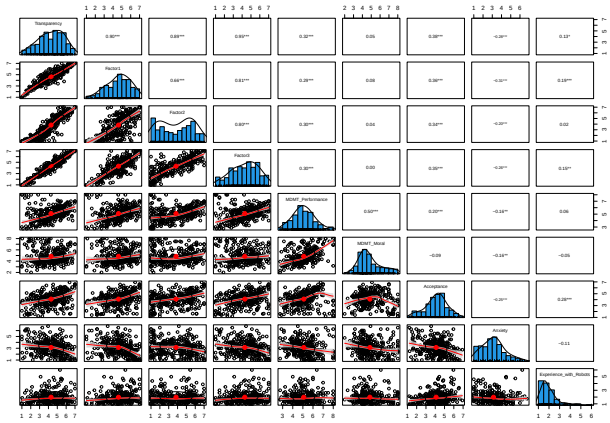
Furthermore, we investigated the effects of the transparency of a robot's behavior on the three factors of the transparency scale. We also examined whether there were differences between language groups regarding these three factor. For Factor 1, the high transparency condition showed significantly higher scores than the low transparency condition ($t(880) = 16.97, p < .001, 95\%$ CI $[1.26, 1.59], d = 1.13$). An ANOVA (see Table 6) showed that both the manipulated transparency and participants' language had significant main effects on Factor 1. Post-hoc tests showed that German participants scored lower than English participants ($t(590) = 6.00, p < .001, 95\%$ CI $[0.48, 0.94], d = 0.49$), who scored lower than Italian participants ($t(590) = -5.46, p < .001, 95\%$ CI $[-0.87, -0.41], d = -0.45$). However, there was no significant difference between German and Italian participants ($t(599) = 0.61, p = 1, 95\%$ CI $[-0.15, 0.28], d = 0.05$), and no interaction effect was found. For Factor 2, the high transparency condition again scored significantly higher than the low transparency condition ($t(872) = 43.78, p < .001, 95\%$ CI $[2.79, 3.06], d = 2.91$). An ANCOVA confirmed the main effect of manipulated transparency on Factor 2 but found no main effect of language. There was, however, a marginally significant interaction between the transparency manipulation and participant language. Lastly, Factor 3 also showed significantly higher scores in the high transparency condition compared to the low transparency condition ($t(880) = 24.36, p < .001, 95\%$ CI $[1.79, 2.11], d = 1.62$). Finally, the ANOVA confirmed the main effect of the experimental manipulation on Factor 3 but found no significant effect of participant language and no interaction effect. Therefore, the transparency of robot behavior influenced the scale, with some variations based on the



(a) The English version of the scale.



(b) The German version of the scale.



(c) The Italian version of the scale.

Figure 5. Correlation matrices between the dependent and control variables of Experiment 2 across the three languages versions.

language background of the participants. Overall, these results provide support for the sentitivity of the scale to a manipulation of the transparency of a robot's behavior.

Finally, how the experimental conditions and participants' language affected other dependent variables was examined (as detailed in Tables S24-S26 of the Supplementary Information). Performance trust was significantly higher in the high transparency condition ($t(857) = 11.06, p < .001, 95\%$ CI $[0.75, 1.08], d = 0.73$). An ANOVA (see Table S24 of the Supplementary Information) showed that both the experimental manipulation and participants' language had significant effects on performance trust. Post-hoc tests (with Bonferroni correction) revealed that performance trust was significantly lower for the German participants than for the English participants, $t(594) = -3.99$, $p < .001$, 95% CI $[-0.64, -0.22]$, $d = -0.33$, and for German participants than for Italian participants, $t(584) = -11.77$, $p < 001$, 95% CI $[-1.32, -0.94]$, $d = -0.96$. Furthermore, performance trust was significantly lower for English participants than for Italian participants, $t(569) = -6.98$, $p < .001$, 95% CI $[-0.90, -0.50]$, $d = -0.57$. There was also a significant interaction between experimental manipulation and language. Moral trust was also higher in the high transparency condition ($t(885) = 5.67, p < .001, 95\%$ CI $[0.37, 0.75], d = 0.38$). An ANCOVA (results presented in Table S25 of the Supplementary Information) revealed that the experimental manipulation and participants' language both had significant effects. Post-hoc tests (with Bonferroni correction) showed that Italian participants had significantly higher moral trust than both English participants,$t(590) = 14.33$, $p < .001$, 95% CI $[1.38, 1.82]$, $d = 1.17$, and German participants, $t(596) = 15.12$, $p < 001$, 95% CI $[1.33, 1.73]$, $d = 1.23$, while there was no significant difference between English and German participants, $t(576) = 0.67$, $p = .501$, 95% CI $[-0.14, -0.29]$, $d = 0.05$. An interaction effect was also observed. Robot acceptance was higher in the high transparency condition ($t(884) = 6.53, p < .001, 95\%$ CI $[0.36, 0.67], d =$

0.44). An ANCOVA indicated that both experimental manipulation and language had significant effects (see Table S26 in the Supplementary Information). Post-hoc tests with Bonferroni correction showed that Italian participants had the highest acceptance, significantly higher than both English, $t(590) = 3.96$, $p < .001$, 95% CI $[0.20, 0.60]$, $d = 0.32$, and German participants, $t(599) = 2.53$, $p = .048$, 95% CI $[0.06, 0.42]$, $d = 0.21$. There was no significant difference between German and English participants, $t(586) = 1.63$, $p = .289$, 95% CI $[-0.03, 0.36]$, $d = 0.13$. No interaction effect was identified by the ANOVA.

Anxiety towards the robot was lower in the high transparency condition ($t(897) = -4.47, p < .001, 95\%$ CI $[-0.56, -0.22], d = -0.30$). An ANCOVA (as illustrated in Table S27 of the Supplementary Information) showed that both experimental manipulation and language had significant effects. Post-hoc tests (with Bonferroni correction) revealed that German participants had significantly less anxiety than English participants, $t(593) = -2.42$, $p = .039$, 95% CI $[-0.48, -0.05]$, $d = -0.20$. No significant differences were found between English and Italian participants, $t(588) = 0.49$, $p = 1$, 95% CI $[-0.16, 0.26]$, $d = 0.04$, or between German and Italian participants, $t(597) = -2.08$, $p = .138$, 95% CI $[-0.42, -0.01]$, $d = -0.17$. There was no interaction effect.

Consistent with Experiment 1, these results confirmed that transparency of a robot's behavior has a positive influence on trust and acceptance of robots. As seen on Figure 5, the dimensions of perceived transparency of a robot's behavior measured by our scale correlated well with trust and acceptance. The consistent pattern of higher transparency corresponding with increased trust and acceptance and decreased anxiety therefore supports the validity of our scale.

Parameter	df	SS	MS	F value	p value	ηp^2	95% CI
Overall Score of the Transparency							
Transparency condition	1	910.30	910.30	907.36	<.001 ***	.50	[0.47, 1.00]
Language	2	19.60	9.80	9.76	<.001***	.02	[0.01, 1.00]
Transparency condition: Language	2	3.60	1.80	1.80	.167	< .01	[0.00, 1.00]
Residuals	895	897.9	1.0				
Factor 1							
Transparency condition	1	456.40	456.40	306.72	<.001 ***	.26	[0.22, 1.00]
Language	2	98.00	49.00	32.92	<.001***	.07	[0.04, 1.00]
Transparency condition: Language	2	0.80	0.40	0.276	.759	< .01	[0.00, 1.00]
Residuals	895	1331.90	1.50				
Factor 2							
Transparency condition	1	1925.70	1925.70	1921.13	<.001 ***	.68	[0.66, 1.00]
Language	2	4.60	2.30	2.28	.103	.01	[0.00, 1.00]
Transparency condition: Language	2	5.90	3.0	2.94	.053	.01	[0.00, 1.00]
Residuals	895	897.10	1.00				
Factor 3							
Transparency condition	1	858.00	858.00	594.44	<.001 ***	.40	[0.36, 1.00]
Language	2	6.30	3.10	2.18	.114	.01	[0.00, 1.00]
Transparency condition: Language	2	6.80	3.40	2.36	.095	.01	[0.00, 1.00]
Residuals	895	1291.90	1.40				

Note: df = Degrees of freedom; SS = Sum of Squares; MS = Mean Squares;
*p< .05; **p< .01; ***p< 0.001;

Table 6. Results of the 2-way ANOVAs conducted on the dependent variables of Experiment 2, with the manipulation of the transparency and the language of participants as independent variables

Discussion

The aim of the present research was to develop and validate a scale to assess the perceived transparency of a robot's behavior. To do so, we designed a scale reflecting a four-factor model of transparency based on state of the art research on transparency. Accordingly, we distinguish the dimensions legibility, predictability, explainability, and meta-understanding. Contrary to our expectations, Experiment 1 showed that the perceived transparency of a robot consists of three factors. Building upon these findings, we propose the Transparency Of RObots Scale (TOROS), based on a three-factor model: Factor 1, *Illegibility* comprises items expressing difficulty in comprehending the robot's functioning, objectives, and processes (e.g., "The robot's overall functioning is a mystery to me", "I cannot comprehend the robot's inner processes"). As such, perceived *Illegibility* does not only depend on how legible the actions of a robot are to observers, but also on how they help them infer the robot's functioning

and goals. That is why *Illegibility* encompassed items that were initially assumed to belong to perceived legibility and perceived meta-understanding. *Illegibility* has been chosen instead of legibility because this factor mostly contained reverse-worded items. However, as concluded in Experiment 2, this overrepresentation of reverse-worded items is probably not a research artefact. Indeed, as humans heavily rely on social norms and expectations to understand and socially interact with other agents[69,70], they are more likely to notice when an agent's behavior violates or fails to meet such norms or expectations and to engage cognitive resources to understand the agent and the meaning of its behavior[69]. Similarly, when an agent's communication and actions are unclear, ambiguous, or uninformative, individuals need to spend more cognitive effort to interpret them[70,71]. Therefore, it is possible that the items with the highest factor loading on Factor 1, the highest item discrimination and the lowest item difficulty were mostly reverse-worded because a lack of legibility was a more salient information than legibility for participants to make an evaluation of a robot's behavior. The Factor 2 *Explainability*, is based on items evaluating perceived quality, clarity, and usefulness of the robot's explanations about its actions and states (e.g., "The robot explains complex tasks in a way that is easy to understand", "The robot provides clear explanations for its actions"). The third factor, *Predictability* represents items assessing the users' ability to anticipate or foresee the robot's future actions based on its current behavior (e.g., "It is easy for me to foresee the robot's future actions", "The robot's behavior is predictable"). While conceptually related[13–15], the three proposed TOROS factors offer a new perspective on measuring the perceived robot transparency, supported by empirical data from three countries. In addition, the results showed that these factors were all sensitive to the experimental manipulation of the explainability, legibility, and predictability of a robot's behavior. Experiment 2 confirmed the factorial structure of the scale in three languages: English, German, and Italian.

Additionally, in both Experiments 1 and 2, the TOROS scale demonstrated a good convergent validity with factors related to transparency, namely trust towards robots and acceptance of robots[16–18]. The experimental manipulation of transparency in both Experiments 1 and 2 had an effect on trust and acceptance and, therefore, confirms that the transparency of a robot's behavior is an important determinant of trust and acceptance in HRI. Notwithstanding, the effects of the experimental manipulation of transparency on the factors of perceived transparency were the most significant and the largest. Hence, TOROS has the potential to provide more accurate estimations of the influence of perceived transparency on trust, acceptance, and other constructs that transparency is supposed to determine.

Taken together, our results confirm that TOROS represents a reliable and valid measure of the perceived transparency of a robot's behavior. Besides, it underlies the discrepancy between theorizing about and implementation of transparency in HRI, and how individuals perceive it. More specifically, the results suggest that any manipulation of transparency mostly influences perceived *Explainability*. Interestingly, we found no interaction effect between manipulated explainability, legibility, and predictability of the robot's behavior on the factors of the scale in Experiment 1, yet all these factors had main effects on the different subscales of the TOROS. This suggests that distinguishing explainability, legibility, and predictability of a robot from perceived transparency of a robot's behaviors is important. Existing theories of transparency and the factorial structure of TOROS are consistent in terms of what constitutes transparency. Nevertheless, the way transparency is implemented in a robot does not result in equivalent perceptions of transparency in a user's mind (e.g., In Experiment 1, making a robot more predictable did not only result in higher perceived predictability). This remains to be confirmed in other scenarios. Future research should delve more into the psychological mechanisms that determine the perceived transparency of a robot or any artificial agent.

Interestingly, in Experiment 2, despite TOROS being administered after participants saw an entire video with a robot reaching its goal, an effect of the experimental manipulation of transparency on *Predictability* factor was still detected. This goes against what can be referred to as the *Valley of the normal*: People tend to find ordinary events to be retrospectively predictable[31]. This is due to the fact that understanding processes of resolved events are "backward-looking": When confronted with unexpected but also unsurprising events, people tend to examine the past in a causal thinking process. This induces them to conclude that such events are self-explanatory and to overestimate their predictability. This phenomenon is known as the hindsight bias[72,73], and could have skewed the answers of participants. Indeed, as perceived transparency was assessed after a robot's behavior was fully resolved, a ceiling effect for *Predictability* could have been observed, and yet was not. Further examination of this phenomenon in future research is required.

TOROS is the first reliable and valid tool to measure perceived transparency of robots' behavior. The scale and its instructions in all three languages are provided in the supplementary material. Researchers can use it to test the influence of selected independent variables on perceived transparency. These independent variables can be features of a robot's behavior (e.g., social cues, structure or content of a robot's utterances), contextual factors that may affect users' ability to interpret the robot behavior (e.g., task difficulty, noisiness of the surrounding environment), individual factors that could affect how people interpret and evaluate the transparency of a robot (e.g., existing attitudes towards robots), or a potential interaction between such variables. TOROS can also be used to assess how perceived transparency determines or correlates with other factors (such as trust or acceptance). Because of the multidimensionality of the scale, researchers can select the dimensions of perceived transparency they specifically intend to explore by using factors of interest as subscales. In addition, the high sensitivity of TOROS to

manipulations of transparency, as shown in Experiments 1 and 2, makes TOROS very relevant as a manipulation check to verify whether an experimental manipulation of transparency was successful. Ultimately, the purpose of TOROS is to provide researchers with a validated tool to limit reliance on self-made and non-tested items, as seen in prior research[23]. As such, TOROS will improve the internal validity and reliability of future empirical results, while improving their comparability by establishing a measurement standard in research on transparency. For TOROS to efficiently and accurately measure perceived transparency of robots' behavior, it must be used as intended. This includes using the same fully labelled 7-point Likert format[42] and the exact same items, as well as randomizing the order of all items[50,74], or at least alternate between positively worded and reverse-worded items to hinder both acquiescence bias and careless responding[40]. If other constructs are measured alongside perceived transparency and the order of the measures does not matter, the items can be mixed together and randomized as well[40,50,74]. Any change of the scale's format, rephrasing of item, or removal of the items should be avoided, or done with strong theoretical or methodological rationales. In case of changes due to justified research needs, the validity and reliability of the scale should be assessed again, at least by conducting a CFA and calculation of internal consistency, to make sure such adjustments did not negatively impact the validity and reliability of the scale. Some researchers may be interested in calculating a composite score for perceived transparency of robots' behavior. As suggested by the CFAs of Experiment 2, perceived transparency of a robot's behavior is not a one-factor construct, and should preferably be addressed as a three-dimensional construct. However, the results of both Experiments 1 and 2 provided support for the sensitivity of the composite score to manipulations of the transparency of a robot's behavior. To calculate a composite score of perceived transparency, the intended coding of the items measuring perceived *Illegibility* should be reversed (i.e., the two items supposed to be reverse-coded should be left unchanged, while the eight items supposed to be left unchanged should be reverse-coded), so its polarity is consistent with perceived *Explainability* and perceived *Predictability*. Once this is achieved, an average of the three subscores of TOROS can be calculated, but should be interpreted with caution, while acknowledging and justifying this scoring as a methodological choice. A CFA based on a 1-Factor model can be considered to assess the validity of this approach in the context of a new study, but is likely to result in fit indices that are as poor as those obtained in Experiment 2.

Despite the promising results of the TOROS scale, the present research does not come without methodological limitations: For instance, a crucial limitation pertains to the measurement invariance results across the different language versions of the scale. Even though configural and metric invariance were achieved, TOROS did not demonstrate full scalar and residual invariance. Besides, there were slight differences in perceived transparency across languages, particularly regarding English and German participants, as well as between English and Italian participants. However, it is important to note that achieving only partial measurement invariance is not uncommon in cross-cultural research, especially when dealing with complex psychological constructs like transparency in HRI. Indeed, partial invariance can still allow for meaningful cross-group comparisons[75]. The strong internal consistency and convergent validity demonstrated across all three language versions suggest that the scale is reliable and valid within each language context. Additionally, the observed differences between languages in terms of perceived transparency were small, and almost no interaction between the language of the participant and the manipulation of transparency was observed. However, further research should investigate cultural factors that may affect how people construe transparency in robots. As such inquiries would require TOROS to be available in the language of the targeted cultures, validations of new translations should be conducted, either before or during these studies. Our current methodology aligns well with established practices in scale development in HRI[25,28,32,76]. Thus, image vignettes and videos of real scenarios provided a strong foundation for scale development. Going beyond such classic approach, we recommend to conduct further validation research in the HRI context, in particular on participants directly interacting with robots (in constrast with participants observing them). We look forward to seeing the TOROS scale put to use to shed further light on the notion of transparency in social robots.

Conclusions

As robots become more sophisticated and prevalent in our society, the need to measure how transparent they are to humans becomes increasingly important. Yet, until now, a standardized method to measure this critical aspect of robot behavior has been lacking. This work addresses this gap by developing and validating the first comprehensive scale to assess the perceived transparency of robotic systems, termed TOROS. Through a rigorous three-stage process involving 1,223 participants, we have created a robust tool encompassing 26 items and comprised of three factors: *Illegibility*, *Explainability*, and *Predictability*. The scale demonstrates high cross-linguistic reliability and validity across English, German, and Italian languages. We believe that the proposed scale can serve as a valuable tool in various HRI experiments to examine the effects of transparency-related aspects on other phenomena. For instance, it can be employed in studies of human-robot collaboration to evaluate how transparency impacts team performance, in learning and adaptation research to track changes in the understanding of the robot as users gain experience with it, and in error recovery and management to assess the effectiveness of error handling strategies. The scale can also be used to investigate how different robot designs and behaviors influence perceived transparency and how this, in turn, affects the overall interaction quality.

Future research should focus on translating the scale into more languages and examining its performance in real HRI scenarios or in interaction with different artificial agents. Follow-up works can use the scale to understand the determinants of perceived transparency in HRI and contribute to a better understanding of the psychological mechanisms at play. This tool opens up new avenues for research and has the potential to significantly enhance our understanding of transparency in robotics, ultimately leading to the development of more effective and user-friendly robotic systems.

Data availability

The datasets generated and analyzed during the current study are available from the corresponding authors upon request.

References

1. Hoffman, G. Anticipation in human-robot interaction. In *AAAI Spring Symposium: It's All in the Timing* (2010).
2. Dragan, A. D., Bauman, S., Forlizzi, J. & Srinivasa, S. S. Effects of robot motion on human-robot collaboration. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*, 51–58 (2015).
3. Theodorou, A., Wortham, R. H. & Bryson, J. J. Designing and implementing transparency for real time inspection of autonomous robots. *Connect. Sci.* **29**, 230–241 (2017).
4. Collins, E. C. Vulnerable users: deceptive robotics. *Connect. Sci.* **29**, 223–229 (2017).
5. Prifti, K. & Fosch-Villaronga, E. Towards experimental standardization for ai governance in the eu. *Comput. Law & Secur. Rev.* **52**, 105959 (2024).
6. Veale, M. & Zuiderveen Borgesius, F. Demystifying the draft eu artificial intelligence act—analysing the good, the bad, and the unclear elements of the proposed approach. *Comput. Law Rev. Int.* **22**, 97–112 (2021).
7. Limongelli, R. Caveat emptor de ai, from black box to glass box: Reviewing consumer protection against ai-enabled manipulation in the european union. *Available at SSRN 4519937* (2023).
8. Miller, C. A. Delegation and transparency: Coordinating interactions so information exchange is no surprise. In *International Conference on Virtual, Augmented and Mixed Reality*, 191–202 (Springer, 2014).
9. Wortham, R. H., Theodorou, A. & Bryson, J. J. Robot transparency: Improving understanding of intelligent behaviour for designers and users. In *Annual Conference Towards Autonomous Robotic Systems*, 274–289 (Springer, 2017).
10. Kim, T. & Hinds, P. Who should i blame? effects of autonomy and transparency on attributions in human-robot interaction. In *ROMAN 2006-The 15th IEEE international symposium on robot and human interactive communication*, 80–85 (IEEE, 2006).
11. Anjomshoae, S., Najjar, A., Calvaresi, D. & Främling, K. Explainable Agents and Robots : Results from a Systematic Literature Review. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 1078–1088 (International Foundation for Autonomous Agents and MultiAgent Systems, 2019). URL https://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-158024.
12. Stange, S. & Kopp, S. Explaining Before or After Acting? How the Timing of Self-Explanations Affects User Perception of Robot Behavior. In Li, H. *et al.* (eds.) *Social Robotics*, 142–153 (Springer International Publishing, Cham, 2021). DOI 10.1007/978-3-030-90525-5_13.
13. Angelopoulos, G., Rossi, A. & Rossi, S. Robot behaviours for transparent human-robot interaction. In *2023 I-RIM Conference*, 186–188 (I-RIM, 2023). DOI 10.5281/zenodo.10722566.
14. Endsley, M. R. From here to autonomy: lessons learned from human–automation research. *Hum. factors* **59**, 5–27 (2017).
15. Alonso, V. & De La Puente, P. System transparency in shared autonomy: A mini review. *Front. neurorobotics* **12**, 83 (2018).
16. Schor, B. G., Norval, C., Charlesworth, E. & Singh, J. Mind the gap: Designers and standards on algorithmic system transparency for users. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–16 (2024).
17. Fischer, K., Weigelin, H. M. & Bodenhagen, L. Increasing trust in human–robot medical interactions: effects of transparency and adaptability. *Paladyn, J. Behav. Robotics* **9**, 95–109 (2018).
18. Aquilino, L., Bisconti, P., Marchetti, A. *et al.* Trust in ai: Transparency, and uncertainty reduction. development of a new theoretical framework. In *CEUR WORKSHOP PROCEEDINGS*, 19–26 (CEUR-WS. org, 2024).
19. Cai, M., Jin, Q., Zhou, J. & Luo, X. How Transparency Shapes the Quality of Human-Robot Interaction: An Examination of Trust, Perception, and Workload. *Int. J. Soc. Robotics* **17**, 1335–1362 (2025). DOI 10.1007/s12369-025-01255-0.

20. Rahwan, I. *et al.* Machine behaviour. *Nat.* **568**, 477–486 (2019). DOI 10.1038/s41586-019-1138-y.

21. Ososky, S., Sanders, T., Jentsch, F., Hancock, P. & Chen, J. Y. Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems. In *Unmanned systems technology XVI*, vol. 9084, 112–123 (SPIE, 2014).

22. Ezenyilimba, A. *et al.* Impact of transparency and explanations on trust and situation awareness in human–robot teams. *J. cognitive engineering decision making* **17**, 75–93 (2023).

23. Schött, S. Y., Amin, R. M. & Butz, A. A literature survey of how to convey transparency in co-located human–robot interaction. *Multimodal Technol. Interact.* **7**, 25 (2023).

24. Claure, H. *et al.* Fairness and transparency in human-robot interaction. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 1244–1246 (IEEE, 2022).

25. Bartneck, C., Kulić, D., Croft, E. & Zoghbi, S. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int. journal social robotics* **1**, 71–81 (2009).

26. Wengefeld, T. *et al.* A laser projection system for robot intention communication and human robot interaction. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 259–265 (IEEE, 2020).

27. Olatunji, S. A. *et al.* Levels of automation and transparency: interaction design considerations in assistive robots for older adults. *IEEE Transactions on Human-Machine Syst.* **51**, 673–683 (2021).

28. Schrum, M. *et al.* Concerning Trends in Likert Scale Usage in Human-robot Interaction: Towards Improving Best Practices. *ACM Transactions on Human-Robot Interact.* **12**, 33:1–33:32 (2023). DOI 10.1145/3572784.

29. Clark, A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* **36**, 181–204 (2013). DOI 10.1017/S0140525X12000477.

30. Vesper, C. *et al.* Joint Action: Mental Representations, Shared Information and General Mechanisms for Coordinating with Others. *Front. Psychol.* **7** (2017). DOI 10.3389/fpsyg.2016.02039.

31. Kahneman, D., Sibony, O. & Sunstein, C. R. *Noise: a flaw in human judgment* (William Collins, London, 2021).

32. Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R. & Young, S. L. Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. *Front. Public Heal.* **6**, 149 (2018). DOI 10.3389/fpubh.2018.00149.

33. Tabachnick, B. G., Fidell, L. S. & Ullman, J. B. *Using multivariate statistics* (Pearson, NY, 2019), seventh edition edn.

34. Lee, M., Ruijten, P., Frank, L. & IJsselsteijn, W. Here's looking at you, robot: The transparency conundrum in hri. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2120–2127 (IEEE, 2023).

35. Angelopoulos, G., Imparato, P., Rossi, A. & Rossi, S. Using theory of mind in explanations for fostering transparency in human-robot interaction. In *International Conference on Social Robotics*, 394–405 (Springer, 2023).

36. Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O. & Weisz, J. D. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–19 (2021).

37. DeVellis, R. F. & Thorpe, C. T. *Scale development: Theory and applications* (Sage publications, 2021).

38. Weijters, B., Cabooter, E. & Schillewaert, N. The effect of rating scale format on response styles: The number of response categories and response category labels. *Int. J. Res. Mark.* **27**, 236–247 (2010). DOI 10.1016/j.ijresmar.2010.02.004.

39. Zhang, X., Noor, R. & Savalei, V. Examining the Effect of Reverse Worded Items on the Factor Structure of the Need for Cognition Scale. *PLoS ONE* **11**, e0157795 (2016). DOI 10.1371/journal.pone.0157795.

40. Weijters, B., Baumgartner, H. & Schillewaert, N. Reversed item bias: An integrative model. *Psychol. Methods* **18**, 320–334 (2013). DOI 10.1037/a0032121.

41. van Sonderen, E., Sanderman, R. & Coyne, J. C. Ineffectiveness of Reverse Wording of Questionnaire Items: Let's Learn from Cows in the Rain. *PLOS ONE* **8**, e68967 (2013). DOI 10.1371/journal.pone.0068967.

42. Taherdoost, H. What Is the Best Response Scale for Survey and Questionnaire Design; Review of Different Lengths of Rating Scale / Attitude Scale / Likert Scale. *Int. J. Acad. Res. Manag.* **8**, 1–10 (2022). URL https://papers.ssrn.com/abstract=4178693.

43. Kaplan, K. J. On the ambivalence-indifference problem in attitude theory and measurement: A suggested modification of the semantic differential technique. *Psychol. Bull.* **77**, 361–372 (1972). DOI 10.1037/h0032590.

44. Colman, A. M., Norris, C. E. & Preston, C. C. Comparing Rating Scales of Different Lengths: Equivalence of Scores from 5-Point and 7-Point Scales. *Psychol. Reports* **80**, 355–362 (1997). DOI 10.2466/pr0.1997.80.2.355.

45. Stapels, J. G. & Eyssel, F. Let's not be indifferent about robots: Neutral ratings on bipolar measures mask ambivalence in attitudes towards robots. *PLOS ONE* **16**, e0244697 (2021). DOI 10.1371/journal.pone.0244697.

46. Chyung, S. Y. Y., Roberts, K., Swanson, I. & Hankinson, A. Evidence-Based Survey Design: The Use of a Midpoint on the Likert Scale. *Perform. Improv.* **56**, 15–23 (2017). DOI 10.1002/pfi.21727.

47. Gilljam, M. & Granbi, D. Should we take don't know for an answer? *Public Opin. Q.* **57**, 348–357 (1993). DOI 10.1086/269380.

48. Ullman, D. & Malle, B. F. Measuring Gains and Losses in Human-Robot Trust: Evidence for Differentiable Components of Trust. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 618–619 (2019). DOI 10.1109/HRI.2019.8673154.

49. Heerink, M., Krose, B., Evers, V. & Wielinga, B. Measuring acceptance of an assistive social robot: a suggested toolkit. In *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*, 528–533 (2009). DOI 10.1109/ROMAN.2009.5326320.

50. Buchanan, E. M. *et al.* Does the delivery matter? Examining randomization at the item level. *Behav.* **45**, 295–316 (2018). DOI 10.1007/s41237-018-0055-y.

51. Reich-Stiebert, N. & Eyssel, F. Learning with Educational Companion Robots? Toward Attitudes on Education Robots, Predictors of Attitudes, and Application Potentials for Education Robots. *Int. J. Soc. Robotics* **7**, 875–888 (2015). DOI 10.1007/s12369-015-0308-9.

52. Hatcher, L. & Stepanski, E. J. *A step-by-step approach to using the SAS system for univariate and multivariate statistics.* (SAS Institute, 1994).

53. Sass, D. A. & Schmitt, T. A. A Comparative Investigation of Rotation Criteria Within Exploratory Factor Analysis. *Multivar. Behav. Res.* **45**, 73–103 (2010). DOI 10.1080/00273170903504810.

54. Fabrigar, L. R., Wegener, D. T., MacCallum, R. C. & Strahan, E. J. Evaluating the use of exploratory factor analysis in psychological research. *Psychol. Methods* **4**, 272–299 (1999). DOI 10.1037/1082-989X.4.3.272.

55. Kaiser, H. F. An index of factorial simplicity. *Psychom.* **39**, 31–36 (1974). DOI 10.1007/BF02291575.

56. Shrestha, N. Factor Analysis as a Tool for Survey Analysis. *Am. J. Appl. Math. Stat.* **9**, 4–11 (2021). DOI 10.12691/ajams-9-1-2.

57. Güvendir, M. A. & Özkan, Y. Ö. Item removal strategies conducted in exploratory factor analysis: A comparative study. *Int. J. Assess. Tools Educ.* **9**, 165–180 (2022).

58. Knekta, E., Runyon, C. & Eddy, S. One size doesn't fit all: Using factor analysis to gather validity evidence when using surveys in your research. *CBE-Life Sci. Educ.* **18**, rm1 (2019).

59. Rahman, I. A. & Al-Emad, N. Structural relationship of leadership qualities with worker's issues for saudi arabia's construction industry. In *MATEC Web of Conferences*, vol. 250, 05002 (EDP Sciences, 2018).

60. Lalor, J. P., Wu, H. & Yu, H. Building an evaluation scale using item response theory. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, vol. 2016, 648 (NIH Public Access, 2016).

61. van der Linden, W. J. (ed.) *Handbook of item response theory*. Chapman & Hall/CRC statistics in the social and behavioral sciences series (CRC Press, Boca Raton, 2016).

62. Feldt, L. S. The Relationship Between the Distribution of Item Difficulties and Test Reliability. *Appl. Meas. Educ.* (1993). DOI 10.1207/s15324818ame0601_3.

63. Mukherjee, P. & Lahiri, S. K. Analysis of Multiple Choice Questions (MCQs): Item and Test Statistics from an assessment in a medical college of Kolkata, West Bengal. *IOSR J. Dental Med. Sci.* **14**, 47–52 (2015).

64. Collins, J. Education techniques for lifelong learning: writing multiple-choice questions for continuing medical education activities and self-assessment modules. *Radiogr. A Rev. Publ. Radiol. Soc. North Am. Inc* **26**, 543–551 (2006). DOI 10.1148/rg.262055145.

65. Date, A. P. *et al.* Item analysis as tool to validate multiple choice question bank in pharmacology. *Int. J. Basic & Clin. Pharmacol.* **8**, 1999–2003 (2019). DOI 10.18203/2319-2003.ijbcp20194106.

66. Woods, C. M. Careless Responding to Reverse-Worded Items: Implications for Confirmatory Factor Analysis. *J. Psychopathol. Behav. Assess.* **28**, 186–191 (2006). DOI 10.1007/s10862-005-9004-7.

67. Fenn, J., Tan, C.-S. & George, S. Development, validation and translation of psychological tests. *BJPsych Adv.* **26**, 306–315 (2020).

68. Vagnetti, R. *et al.* Instruments for measuring psychological dimensions in human-robot interaction: Systematic review of psychometric properties. *J. medical Internet research* **26**, e55597 (2024).

69. Mooney, A. Co-operation, violations and making sense. *J. Pragmat.* **36**, 899–920 (2004). DOI 10.1016/j.pragma.2003.10.006.

70. FeldmanHall, O. & Shenhav, A. Resolving uncertainty in a social world. *Nat. Hum. Behav.* **3**, 426–435 (2019). DOI 10.1038/s41562-019-0590-x.

71. Derex, M. & Boyd, R. Social information can potentiate understanding despite inhibiting cognitive effort. *Sci. Reports* **8**, 9980 (2018). DOI 10.1038/s41598-018-28306-z.

72. Fischhoff, B. Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *J. Exp. Psychol. Hum. Percept. Perform.* **1**, 288–299 (1975). DOI 10.1037/0096-1523.1.3.288.

73. Roese, N. J. & Vohs, K. D. Hindsight Bias. *Perspectives on Psychol. Sci.* **7**, 411–426 (2012). DOI 10.1177/1745691612454303.

74. Knowles, E. S. *et al.* Order Effects within Personality Measures. In Schwarz, N. & Sudman, S. (eds.) *Context Effects in Social and Psychological Research*, 221–236 (Springer, New York, NY, 1992). DOI 10.1007/978-1-4612-2848-6_15.

75. Robitzsch, A. & Lüdtke, O. Why full, partial, or approximate measurement invariance are not a prerequisite for meaningful and valid group comparisons. *Struct. Equ. Model. A Multidiscip. J.* **30**, 859–870 (2023).

76. Carpinella, C. M., Wyman, A. B., Perez, M. A. & Stroessner, S. J. The robotic social attributes scale (rosas) development and validation. In *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction*, 254–262 (2017).

Acknowledgements

The authors thank Francesco Vigni and Lena Schubert for translating the scale into Italian and German language, respectively. This work has been supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955778.

Author contributions statement

G.A., D.L., and F.E. developed the scale. G.A. and D.L. designed the experiments. R.W., A.R., S.R., and F.E. gave feedback on and supported the development of the scale and experimental material. G.A. and D.L. conducted the experiments and analyzed the results. R.W. translated the scale into German. A.R. translated the scale into Italian. G.A. and D.L. wrote the initial draft. R.W., A.R., S.R. and F.E. gave feedback on and refined subsequent drafts of the manuscript. All co-authors have read and agreed to the submitted version of the manuscript.

Competing interests

The authors declare no competing interests.

Measuring Transparency in Intelligent Robots

Georgios Angelopoulos[1,*,+], Dimitri Lacroix[2,,+], Ricarda Wullenkord[2], Alessandra Rossi[1], Silvia Rossi[1], and Friederike Eyssel[2]**

[1]Interdepartmental Center for Advances in Robotic Surgery - ICAROS, University of Naples Federico II, Naples, 80131, Italy
[2]Center for Cognitive Interaction Technology - CITEC, Bielefeld University, Bielefeld, 33619, Germany
[*]georgios.angelopoulos@unina.it
[**]dimitri.lacroix@uni-bielefeld.de
[+]these authors contributed equally to this work

Supplementary Information

Item Generation

ID	Subscale	Item
TR1	**Legibility**	It is clear to me what the robot does.
TR2		I can easily understand the robot's actions.
TR3		I can quickly grasp the task the robot performs.
TR4		I find the robot's behavior easy to understand.
TR5		I feel informed about the robot's activities.
TR6		I do not understand the robot's actions.
TR7		The robot's behavior is confusing to me
TR8		I am unsure what the robot does.
TR9		The robot's actions are obvious.
TR10		The robot provides information about its actions.
TR11		The robot's cues provide relevant information about what it does.
TR12		The robot's behavior makes sense.
TR13		The robot's behavior is legible
TR14		The robot's actions are hard to follow.
TR15		The robot's behavior is difficult to read.
TR16		It is impossible to know what the robot does.
TR17	**Predictability**	I can predict what the robot will do next.
TR18		The robot's next steps are clear to me.
TR19		I feel confident in predicting the robot's next moves.
TR20		It is easy for me to foresee the robot's future actions.
TR21		I see the pattern in the robot's behavior.
TR22		I cannot anticipate what the robot's goal is.
TR23		It is difficult for me to tell what the robot will do next.
TR24		I feel that the robot is inconsistent.
TR25		The robot's actions are consistent.
TR26		The robot's behavior is predictable.
TR27		The robot provides cues that help predict its next actions.
TR28		It is easy to anticipate what will follow the robot's behavior.
TR29		The robot's past actions help to predict its future behavior.
TR30		The robot's behavior does not help predict what it will do next.
TR31		The robot's actions are unpredictable.
TR32		The robot acts in a way that is random.

ID	Subscale	Item
TR33	**Explainability**	I find the robot's explanations informative.
TR34		I understand why the robot performs its actions.
TR35		The robot's explanations make its actions clear to me.
TR36		I feel like the robot's explanations are useful.
TR37		I feel like the robot's explanations are necessary to understand its actions.
TR38		I find the robot's explanations incoherent.
TR39		The robot's explanations do not make sense to me.
TR40		I cannot explain the robot's behavior.
TR41		The robot provides clear explanations for its actions.
TR42		The robot's behavior is explainable
TR43		The robot provides detailed explanations of its actions.
TR44		The robot's explanations for its actions are straightforward.
TR45		The robot explains complex tasks in a way that is easy to understand.
TR46		The robot's explanations are overly detailed.
TR47		The robot provides unnecessary explanations.
TR48		The robot does not provide enough explanations.
TR49	**Meta Understanding**	The robot's behavior helps me understand its objectives.
TR50		I have a clear understanding of how the robot operates in general.
TR51		I feel like the robot helps me understand its inner processes.
TR52		I am confident in understanding of the robot's overall behavior.
TR53		The robot provides enough cues for me to understand its overall functioning.
TR54		I cannot comprehend the robot's inner processes.
TR55		The robot's overall functioning is a mystery to me.
TR56		I am confused about the robot's general objectives.
TR57		The robot conveys its overall state effectively.
TR58		The robot ensures that its users are well-informed about its activities.
TR59		The robot's functioning is transparent.
TR60		The clarity of the robot's overall functioning eliminates any doubts about it.
TR61		The robot's inner processes are obvious.
TR62		The robot does not provide enough information about its overall objectives.
TR63		It is hard to make sense of the robot's general functioning.
TR64		It is difficult to get a clear picture of the robot's overall operations.

Table S1. The 64 initial items of the scale.

For ease of reference, each item in the scale was assigned a unique identifier, ranging from TR1 to TR64, where 'TR' stands for 'Transparency' and the number indicates the item's position in the scale.

Pre-test

Before Experiment 1, we conducted a *pretest* to identify a hypothetical everyday life scenario featuring HRI that would effectively discriminate transparency. This way, we maximize the scale's sensitivity and effectiveness for measuring transparency. Consequently, we developed four image vignettes, each depicting a robot performing a task in an everyday life scenario (i.e., Scenario 1: A robot heading towards a charging station to refill its battery; Scenario 2: A robot taking an apple core to throw it in a bin; Scenario 3: Arobot switching the light on; and Scenario 4: A robot taking a guest's coat to put in on a rack). The pretest was implemented on Qualtrics and employed a $4 \times 2 \times 2 \times 2$ mixed between-within-subjects design that manipulated explainability, legibility, and predictability of a robot's behavior as a between-subjects factor and treated the context of the scenario as a within-subject factor, resulting in 32 images vignettes total. Transparency was assessed with three items that evaluated the explainability, legibility, and predictability of the respective robot's behavior. Additionally, one item evaluated the understandability of the picture vignettes, and an attention check was implemented at the study's beginning. Furthermore, demographic questions (i.e., age, gender, education, self-assessed English language proficiency) and prior experience with robots. Only complete datasets from participants over 18 years of age and with a self-declared English proficiency at A2 level (Elementary) and above were included. The pretest was designed to run for two weeks or to be discontinued if a sample size of 160 participants was reached before the end of this period. The two-week duration was chosen to balance the need for timely data collection and to allow for sufficient time to recruit participants. Participants for the pretest were recruited via mailing lists, social media, and snowball sampling. After analyzing the pretest, one scenario with eight image vignettes was selected for further use, based on the greatest contrast of mean between high and low transparency. The pretest's pre-registration details are available at https://aspredicted.org/ZTT_STV.

As outlined in our pre-registered criteria, the pretest stopped after two weeks, allowing us to recruit 97 participants through social media and snowball sampling. Following our pre-registered exclusion criteria, 31 participants were disqualified; 24 failed to complete the study, 6 were excluded for not passing both attention checks, and 1 was removed upon request for data removal. Thus, the final sample comprised $N = 66$ participants.

Pretest results revealed that Scenario 1 (i.e., a robot heading towards a charging station to refill its battery) elicited the most substantial variation in responses across all dependent variables, particularly in the measures of explainability and predictability, as depicted in Table S2. This variation, as evident in descriptive statistics such as means and standard deviations, indicates that Scenario 1 elicited a greater range in participant responses, making it the most effective in distinguishing among the different conditions tested. For example, for Scenario 1, the mean scores for transparency were lower when explainability, legibility, and predictability were absent ($M = 3.500$, $SD = 1.225$), which is consistent with expectations. Conversely, higher mean scores for transparency were observed in Scenario 1 when all three factors were present ($M = 6.200$, $SD = 1.135$), demonstrating the scenario's sensitivity to the manipulations. Additionally, we observed that three participants in the low transparency condition (low explainability, low predictability, and low explainability, respectively) reported being able to anticipate the robot's objectives, as these were accentuated through color highlighting, thereby enhancing the robot's predictability and the salience of its goal. Based on these findings, we selected Scenario 1. The sole modification implemented in Scenario 1 involved extending the color application to the entire environment rather than restricting it to specific objects. This modification was hypothesized to amplify the distinction between the conditions within Scenario 1 without altering the relative effectiveness of the manipulations across different scenarios. Therefore, at this point, it was decided to proceed with Experiment 1.

ID	Scenario			Dependent Variables							
	Manipulation			Explainability		Predictability		Legibility		Transparency	
	Explainability	Predictability	Legibility	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev
1	NO	NO	NO	4.17	1.329	3.83	1.472	4.17	1.329	3.50	1.225
	NO	NO	YES	5.90	1.853	5.00	1.700	5.50	1.900	5.10	1.729
	NO	YES	NO	5.38	2.200	5.25	2.435	5.63	2.326	5.50	2.262
	YES	NO	NO	4.33	2.338	3.67	1.966	4.17	2.483	4.00	2.280
	NO	YES	YES	6.00	2.236	5.29	2.430	5.86	2.193	5.14	2.116
	YES	YES	NO	7.00	.000	7.00	.000	6.67	.707	6.78	.667
	YES	NO	YES	6.00	1.155	6.10	1.101	5.80	1.317	5.60	1.506
	YES	YES	YES	6.60	.966	6.00	1.414	6.30	.949	6.20	1.135
2	NO	NO	NO	4.83	2.137	4.83	2.137	5.17	1.835	5.00	2.000
	NO	NO	YES	6.30	.949	5.90	1.449	5.50	1.354	6.00	1.155
	NO	YES	NO	5.63	2.326	5.63	2.200	5.75	2.053	5.25	2.375
	YES	NO	NO	4.33	2.160	4.67	1.211	4.83	2.401	4.83	2.280
	NO	YES	YES	6.14	.900	5.71	1.496	6.14	1.069	5.71	1.380
	YES	YES	NO	6.33	1.000	6.44	.882	6.33	.866	6.11	.928
	YES	NO	YES	5.30	1.418	5.00	.943	5.40	1.265	4.80	1.506
	YES	YES	YES	6.00	1.247	5.70	1.059	5.80	1.398	5.50	1.269
3	NO	NO	NO	4.83	1.941	4.83	1.941	4.67	1.751	5.00	2.098
	NO	NO	YES	4.10	2.378	4.10	2.132	4.40	2.221	3.70	2.058
	NO	YES	NO	5.37	2.264	5.38	2.722	5.88	2.031	5.50	2.070
	YES	NO	NO	4.83	1.722	4.67	1.506	5.17	1.722	5.17	1.941
	NO	YES	YES	5.43	1.813	4.57	2.299	5.71	1.799	4.43	2.370
	YES	YES	NO	5.11	1.691	5.44	2.128	5.22	2.048	5.00	1.803
	YES	NO	YES	5.80	1.398	5.30	1.767	5.50	1.780	5.70	1.494
	YES	YES	YES	5.60	1.506	5.30	1.636	5.60	1.075	5.50	1.179
4	NO	NO	NO	5.17	1.835	5.17	2.137	5.17	.983	5.00	2.000
	NO	NO	YES	6.00	1.155	5.70	1.252	5.20	1.398	5.70	1.160
	NO	YES	NO	5.00	2.070	4.13	2.167	4.00	2.268	4.13	1.885
	YES	NO	NO	4.33	1.966	3.83	1.835	5.33	1.633	5.00	1.673
	NO	YES	YES	6.57	.535	6.43	.787	5.71	1.380	5.86	1.345
	YES	YES	NO	6.11	1.691	4.11	1.965	5.44	1.878	5.56	1.236
	YES	NO	YES	4.60	1.713	4.60	1.350	4.70	1.703	4.70	1.494
	YES	YES	YES	4.80	1.814	4.50	1.958	4.60	1.838	4.80	1.476

Table S2. The results of pre-test.

Experiment 1

Conditions (C1-C8)	No. of Participants	Age		Gender			Experience with robots	
		M	SD	Male	Female	Other	M	SD
C1 - **High Explainability**, **High Legibility**, **High Predictability**	40	27.88	7.95	17	22	1	2.04	1.05
C2 - **High Explainability**, **High Legibility**, Low Predictability	38	31.37	10.24	15	22	1	2.17	1.19
C3 - **High Explainability**, Low Legibility, **High Predictability**	43	28.63	7.86	15	28	0	1.97	1.10
C4 - Low Explainability, **High Legibility**, **High Predictability**	39	30.69	7.88	21	18	0	1.88	1.00
C5 - **High Explainability**, Low Legibility, Low Predictability	38	32.61	11.50	15	23	0	2.14	1.23
C6 - Low Explainability, **High Legibility**, Low Predictability	41	28.73	11.35	19	20	2	2.30	1.27
C7 - Low Explainability, Low Legibility, **High Predictability**	43	30.53	9.72	16	27	0	2.05	1.17
C8 - Low Explainability, Low Legibility, Low Predictability	40	28.4	7.03	17	22	1	2.08	1.03
Total	**322**	**29.82**	**9.34**	**135**	**182**	**5**	**2.08**	**1.13**

Table S3. Demographic information across the eight conditions.

Initial item removals, including TR37, TR46, and TR47, were excluded due to low loadings in the initial factor structure matrix. Following this removal, a new EFA indicated the need to exclude TR4, TR6, TR7, TR11, TR15, TR21, TR22, TR29, TR34, TR39, TR49, TR52, TR59, TR60 due to insufficient loadings, and TR62 for loading on multiple factors. Further iterations excluded TR3, TR25, TR42, and TR61 due to low loadings, and TR2 and TR12 for constituting underrepresented factors. The fourth iteration led to removing TR14 and TR24 due to inadequate loadings, TR31 for multiple factor loadings, and TR32 and TR38 for being in underrepresented factors. The fifth iteration led to the removal of TR48 for loading on various factors and TR51 and TR53 due to low loadings. Ultimately, a final iteration confirmed that no further items required removal.

Items	Initial	Extraction
It is clear to me what the robot does.	.790	.744
I feel informed about the robot's activities.	.765	.705
I am unsure what the robot does.	.719	.617
The robot's actions are obvious.	.719	.604
The robot provides information about its actions.	.781	.738
It is impossible to know what the robot does.	.687	.594
I can predict what the robot will do next.	.747	.700
The robot's next steps are clear to me.	.760	.707
I feel confident in predicting the robot's next moves.	.741	.685
It is easy for me to foresee the robot's future actions.	.699	.630
It is difficult for me to tell what the robot will do next.	.715	.681
The robot's behavior is predictable.	.760	.699
The robot provides cues that help predict its next actions.	.672	.549
It is easy to anticipate what will follow the robot's behavior.	.686	.633
The robot's behavior does not help predict what it will do next.	.617	.523
I find the robot's explanations informative.	.796	.761
The robot's explanations make its actions clear to me.	.817	.764
I feel like the robot's explanations are useful.	.746	.698
I cannot explain the robot's behavior.	.739	.642
The robot provides clear explanations for its actions.	.753	.695
The robot provides detailed explanations of its actions.	.705	.636
The robot's explanations for its actions are straightforward.	.800	.736
The robot explains complex tasks in a way that is easy to understand.	.735	.690
I have a clear understanding of how the robot operates in general.	.688	.637
I cannot comprehend the robot's inner processes.	.603	.544
The robot's overall functioning is a mystery to me.	.650	.582
I am confused about the robot's general objectives.	.711	.601
The robot conveys its overall state effectively.	.647	.569
The robot ensures that its users are well-informed about its activities.	.760	.711
It is hard to make sense of the robot's general functioning.	.635	.567
It is difficult to get a clear picture of the robot's overall operations.	.636	.563
I can easily understand the robot's actions.	.790	.741

Items	Initial	Extraction
I can quickly grasp the task the robot performs.	.734	.637
I find the robot's behavior easy to understand.	.789	.734
I do not understand the robot's actions.	.752	.673
The robot's behavior is confusing to me.	.730	.661
The robot's cues provide relevant information about what it does.	.679	.560
The robot's behavior makes sense.	.755	.703
The robot's behavior is legible.	.692	.601
The robot's actions are hard to follow.	.693	.598
The robot's behavior is difficult to read.	.746	.648
I see the pattern in the robot's behavior.	.646	.523
I cannot anticipate what the robot's goal is.	.654	.561
I feel that the robot is inconsistent.	.553	.402
The robot's actions are consistent.	.662	.547
The robot's past actions help to predict its future behavior.	.467	.325
The robot's actions are unpredictable.	.748	.674
The robot acts in a way that is random.	.609	.471
I understand why the robot performs its actions.	.775	.680
I feel like the robot's explanations are necessary to understand its actions.	.405	.247
I find the robot's explanations incoherent.	.575	.520
The robot's explanations do not make sense to me.	.588	.518
The robot's behavior is explainable.	.688	.618
The robot's explanations are overly detailed.	.428	.240
The robot provides unnecessary explanations.	.332	.195
The robot does not provide enough explanations.	.700	.658
The robot's behavior helps me understand its objectives.	.788	.734
I feel like the robot helps me understand its inner processes.	.677	.597
I am confident in understanding of the robot's overall behavior.	.769	.698
The robot provides enough cues for me to understand its overall functioning.	.753	.689
The robot's functioning is transparent.	.723	.640
The clarity of the robot's overall functioning eliminates any doubts about it.	.701	.635
The robot's inner processes are obvious.	.675	.504
The robot does not provide enough information about its overall objectives.	.686	.623

Table S4. The communalities of the 64 initial items.

Item	Factor				
	1	2	3	4	5
I find the robot's behavior easy to understand.	**.837**	-.013	.095	.058	.144
The robot's behavior helps me understand its objectives.	**.834**	.106	.094	.035	.129
I can easily understand the robot's actions.	**.829**	-.013	.112	.097	.177
I am confident in understanding of the robot's overall behavior.	**.816**	.059	.033	-.161	.029
The robot's explanations make its actions clear to me.	**.813**	.267	-.038	.138	-.106
It is clear to me what the robot does.	**.811**	.042	-.095	-.174	.213
The robot provides enough cues for me to understand its overall functioning.	**.808**	.164	-.018	-.096	-.024
I feel informed about the robot's activities.	**.799**	.227	-.109	.046	.012
The robot's explanations for its actions are straightforward.	**.798**	.221	-.016	.161	-.155
The robot's next steps are clear to me.	**.793**	-.005	.278	-.019	.016
I understand why the robot performs its actions.	**.789**	-.046	.038	.076	.219
The robot's functioning is transparent.	**.781**	.085	-.062	-.103	.095
The robot ensures that its users are well-informed about its activities.	**.772**	.312	-.114	.040	-.059
The robot's behavior makes sense.	**.771**	-.079	.072	.197	.243
The robot provides information about its actions.	**.766**	.351	-.043	.106	-.125
I find the robot's explanations informative.	**.765**	.358	-.096	.158	-.115
The robot provides clear explanations for its actions.	**.763**	.268	-.055	.095	-.168
I can quickly grasp the task the robot performs.	**.763**	.022	.067	-.093	.204
The robot's behavior is difficult to read.	**-.762**	.242	.074	-.034	-.041
The clarity of the robot's overall functioning eliminates any doubts about it.	**.761**	.166	-.038	-.157	.043
I do not understand the robot's actions.	**-.755**	.248	.116	-.134	-.096
The robot's behavior is legible.	**.750**	.030	.098	.134	.100
The robot does not provide enough explanations.	**-.747**	-.014	.280	.016	.145
The robot's behavior is predictable.	**.744**	-.086	.362	-.053	-.068
I feel confident in predicting the robot's next moves.	**.741**	-.092	.342	-.081	-.067
The robot's actions are obvious.	**.737**	.003	.239	.053	.017
The robot's behavior is explainable.	**.737**	-.031	.060	.143	.222
I have a clear understanding of how the robot operates in general.	**.727**	.132	-.071	-.284	.070
The robot's cues provide relevant information about what it does.	**.726**	.160	-.007	-.003	.086
The robot's behavior is confusing to me.	**-.726**	.284	.110	-.088	-.185
I feel like the robot helps me understand its inner processes.	**.721**	.235	-.072	-.115	-.063
The robot does not provide enough information about its overall objectives.	**-.721**	.032	.263	.107	.151
I am unsure what the robot does.	**-.719**	.170	.223	.112	-.090
I cannot explain the robot's behavior.	**-.717**	.347	.069	.040	-.031
The robot explains complex tasks in a way that is easy to understand.	**.717**	.401	-.116	-.039	-.003
It is impossible to know what the robot does.	**-.715**	.228	.174	.001	-.024
I can predict what the robot will do next.	**.712**	-.093	.411	-.069	-.107
I am confused about the robot's general objectives.	**-.709**	.092	.244	.172	-.021
I see the pattern in the robot's behavior.	**.705**	.038	.137	.035	.070
I cannot anticipate what the robot's goal is.	**-.704**	.218	-.022	.122	.056
It is easy to anticipate what will follow the robot's behavior.	**.704**	-.063	.293	-.136	-.171
The robot conveys its overall state effectively.	**.702**	.189	-.072	.185	-.031
The robot provides cues that help predict its next actions.	**.702**	.095	.203	.029	-.073
It is difficult for me to tell what the robot will do next.	**-.700**	.296	-.158	.111	.256
The robot's actions are consistent.	**.699**	-.016	.203	.107	.077
I feel like the robot's explanations are useful.	**.696**	.376	-.136	.197	-.118
The robot's actions are hard to follow.	**-.690**	.283	-.030	-.200	.006
The robot's actions are unpredictable.	**-.688**	.366	-.119	-.025	.228
The robot provides detailed explanations of its actions.	**.677**	.402	-.036	-.075	-.097
The robot's overall functioning is a mystery to me.	**-.673**	.209	.237	.164	-.044
It is easy for me to foresee the robot's future actions.	**.673**	-.084	.388	-.112	-.085
The robot's explanations do not make sense to me.	**-.659**	.116	.161	-.207	.022
It is difficult to get a clear picture of the robot's overall operations.	**-.658**	.124	.219	.256	.018
It is hard to make sense of the robot's general functioning.	**-.655**	.211	.274	.131	.021
The robot's inner processes are obvious.	**.652**	.103	.030	-.258	-.011
The robot's behavior does not help predict what it will do next.	**-.652**	.229	-.130	.001	.166
I cannot comprehend the robot's inner processes.	**-.629**	.220	.222	.199	.107
The robot acts in a way that is random.	**-.536**	.408	.063	-.112	.026
The robot's past actions help to predict its future behavior.	**.530**	.071	.129	.065	.135
I feel that the robot is inconsistent.	**-.521**	.326	.115	-.093	.056
I find the robot's explanations incoherent.	**-.509**	.318	.186	-.336	.111
The robot's explanations are overly detailed.	.150	.407	-.035	-.206	.089
I feel like the robot's explanations are necessary to understand its actions.	.283	.296	-.025	.280	-.013
The robot provides unnecessary explanations.	-.197	.262	.185	-.199	.118

Table S5. The factor structure matrix with the 64 initial items of the scale.

Item	Factor				
	1	2	3	4	5
The robot's behavior makes sense.	**.651**	-.122	-.004	.115	-.106
The robot's behavior is explainable.	**.585**	-.064	-.041	.131	-.103
I understand why the robot performs its actions.	**.573**	-.038	-.137	.103	-.124
The robot's behavior is confusing to me.	**-.535**	.246	.287	.043	.019
I can easily understand the robot's actions.	**.534**	-.029	-.056	.157	-.236
I can quickly grasp the task the robot performs.	**.480**	.127	-.236	.049	-.197
I find the robot's behavior easy to understand.	**.475**	-.020	-.106	.161	-.257
It is clear to me what the robot does.	**.460**	.146	-.456	.083	-.048
I do not understand the robot's actions.	**-.426**	.296	.250	-.083	.064
The robot's behavior helps me understand its objectives.	**.424**	.063	-.100	.261	-.244
The robot's behavior is legible.	**.405**	-.058	-.003	.246	-.229
The robot's actions are consistent.	.368	-.039	.070	.149	-.367
The robot's past actions help to predict its future behavior.	.367	.054	.033	.137	-.177
I see the pattern in the robot's behavior.	.321	.020	-.040	.181	-.310
The robot's behavior is difficult to read.	-.319	.230	.300	-.049	.186
I find the robot's explanations incoherent.	-.130	**.566**	.075	-.225	.013
The robot provides unnecessary explanations.	.022	**.417**	.095	-.095	-.053
The robot's explanations are overly detailed.	.031	**.405**	-.165	.233	.090
The robot acts in a way that is random.	-.214	.393	.194	.081	.190
I feel that the robot is inconsistent.	-.142	.357	.233	-.011	.135
The robot's actions are hard to follow.	-.305	.353	.052	-.085	.269
The robot's explanations do not make sense to me.	-.227	.320	.165	-.304	.015
It is difficult to get a clear picture of the robot's overall operations.	-.075	.016	**.623**	-.029	.101
I cannot comprehend the robot's inner processes.	.026	.148	**.585**	-.029	.159
The robot's overall functioning is a mystery to me.	-.209	.115	**.583**	.015	.030
I am confused about the robot's general objectives.	-.164	.052	**.579**	-.110	.025
It is hard to make sense of the robot's general functioning.	-.118	.175	**.579**	-.052	.022
I am unsure what the robot does.	-.301	.109	**.530**	-.031	.000
The robot does not provide enough information about its overall objectives.	.068	.135	**.523**	-.321	.100
I have a clear understanding of how the robot operates in general.	.183	.235	**-.485**	.149	-.168
The robot does not provide enough explanations.	.031	.177	**.455**	-.428	.044
It is impossible to know what the robot does.	-.257	.232	**.400**	-.071	.080
I cannot explain the robot's behavior.	-.286	.241	.371	.099	.235
The robot's inner processes are obvious.	.070	.204	-.357	.131	-.316
The clarity of the robot's overall functioning eliminates any doubts about it.	.194	.164	-.346	.273	-.186
The robot's functioning is transparent.	.302	.088	-.345	.218	-.129
I am confident in understanding of the robot's overall behavior.	.215	.115	-.327	.177	-.321
I feel like the robot's explanations are useful.	.017	-.047	-.052	**.791**	-.008
I find the robot's explanations informative.	.035	-.022	-.074	**.757**	-.091
The robot provides information about its actions.	.012	.024	-.076	**.705**	-.176
The robot provides clear explanations for its actions.	-.041	-.039	-.112	**.657**	-.217
The robot's explanations make its actions clear to me.	.077	-.044	-.075	**.653**	-.192
The robot's explanations for its actions are straightforward.	.020	-.099	-.041	**.645**	-.254
The robot ensures that its users are well-informed about its activities.	.084	.050	-.207	**.617**	-.080
The robot explains complex tasks in a way that is easy to understand.	.107	.184	-.245	**.595**	-.022
The robot provides detailed explanations of its actions.	-.041	.201	-.193	**.591**	-.180
The robot conveys its overall state effectively.	.177	-.098	-.052	**.544**	-.067
I feel like the robot's explanations are necessary to understand its actions.	.102	-.050	.211	**.538**	.091
I feel informed about the robot's activities.	.211	.021	-.225	**.509**	-.061
I feel like the robot helps me understand its inner processes.	.030	.126	-.309	**.426**	-.183
The robot provides enough cues for me to understand its overall functioning.	.134	.096	-.287	.358	-.254
The robot's cues provide relevant information about what it does.	.301	.081	-.177	.313	-.128
I can predict what the robot will do next.	.087	.030	.080	.018	**-.816**
It is easy for me to foresee the robot's future actions.	.089	.070	.035	-.019	**-.774**
It is easy to anticipate what will follow the robot's behavior.	-.047	.034	-.070	.083	**-.746**
The robot's behavior is predictable.	.151	.021	.043	.037	**-.736**
I feel confident in predicting the robot's next moves.	.141	.032	.001	.020	**-.723**
It is difficult for me to tell what the robot will do next.	.135	.201	.212	.004	**.707**
The robot's actions are unpredictable.	.047	.339	.140	-.010	**.613**
The robot's next steps are clear to me.	.273	.049	.002	.121	**-.565**
The robot's behavior does not help predict what it will do next.	-.001	.208	.115	-.060	**.541**
The robot provides cues that help predict its next actions.	.111	.024	.030	.293	**-.478**
The robot's actions are obvious.	.277	-.005	.048	.169	**-.476**
I cannot anticipate what the robot's goal is.	-.120	.110	.329	.018	.391

Table S6. The factor pattern matrix with the 64 initial items of the scale.

Item	1	2	3	4	5
The robot's explanations make its actions clear to me.	**.831**	.286	.049	-.143	-.050
The robot's explanations for its actions are straightforward.	**.817**	.244	.067	-.169	-.066
I feel informed about the robot's activities.	**.805**	.205	-.084	.058	.054
The robot's next steps are clear to me.	**.797**	-.125	.246	.050	.100
It is clear to me what the robot does.	**.796**	-.026	-.162	.165	.154
The robot provides information about its actions.	**.789**	.331	.032	-.045	-.008
I find the robot's explanations informative.	**.789**	.362	-.004	-.076	-.002
The robot provides clear explanations for its actions.	**.785**	.273	.031	-.054	.014
The robot ensures that its users are well-informed about its activities.	**.781**	.311	-.054	.046	-.018
I feel confident in predicting the robot's next moves.	**.744**	-.227	.279	.051	-.084
The robot explains complex tasks in a way that is easy to understand.	**.739**	.323	-.072	.150	.113
I have a clear understanding of how the robot operates in general.	**.738**	-.005	-.163	.264	-.148
The robot's behavior is predictable.	**.733**	-.181	.310	.046	-.044
The robot's actions are obvious.	**.724**	-.043	.214	-.031	.032
I can predict what the robot will do next.	**.723**	-.252	.360	.030	-.006
I feel like the robot's explanations are useful.	**.720**	.422	-.042	-.125	-.060
It is easy to anticipate what will follow the robot's behavior.	**.717**	-.202	.256	.018	-.084
It is difficult for me to tell what the robot will do next.	**-.709**	.363	-.130	.134	.046
The robot provides cues that help predict its next actions.	**.708**	.016	.247	.081	.218
I am unsure what the robot does.	**-.706**	.197	.282	.022	-.231
The robot conveys its overall state effectively.	**.705**	.222	-.007	-.090	.058
I cannot explain the robot's behavior.	**-.705**	.359	.161	.162	-.067
The robot provides detailed explanations of its actions.	**.703**	.307	.025	.159	-.112
I am confused about the robot's general objectives.	**-.702**	.134	.317	.000	.204
It is impossible to know what the robot does.	**-.697**	.201	.224	.097	-.021
It is easy for me to foresee the robot's future actions.	**.678**	-.228	.331	.073	-.044
It is difficult to get a clear picture of the robot's overall operations.	**-.669**	.184	.305	-.117	.048
The robot's overall functioning is a mystery to me.	**-.664**	.229	.333	-.020	.085
The robot's behavior does not help predict what it will do next.	**-.643**	.259	-.092	.115	.008
It is hard to make sense of the robot's general functioning.	**-.636**	.206	.335	.084	-.011
I cannot comprehend the robot's inner processes.	**-.619**	.224	.262	-.025	-.025

Table S7. The factor structure matrix after removing items with low factor loadings and items that loaded on multiple factors.

Item	Factor				
	1	2	3	4	5
I feel like the robot's explanations are useful.	**.960**	.128	.021	-.056	-.026
I find the robot's explanations informative.	**.865**	.009	-.012	.007	.021
The robot's explanations make its actions clear to me.	**.847**	-.104	.002	-.090	.001
The robot's explanations for its actions are straightforward.	**.810**	-.142	.006	-.130	-.004
The robot provides information about its actions.	**.804**	-.069	.009	.030	.010
The robot provides clear explanations for its actions.	**.732**	-.098	-.033	.021	.036
The robot ensures that its users are well-informed about its activities.	**.709**	-.017	-.114	.120	-.033
The robot conveys its overall state effectively.	**.650**	-.037	-.091	-.009	.087
The robot explains complex tasks in a way that is easy to understand.	**.604**	.014	-.146	.273	.061
The robot provides detailed explanations of its actions.	**.604**	-.152	.010	.191	-.152
I feel informed about the robot's activities.	**.563**	-.052	-.243	.145	.035
I can predict what the robot will do next.	-.022	**-.873**	.026	.003	.028
It is easy for me to foresee the robot's future actions.	-.037	**-.832**	.024	.036	-.025
I feel confident in predicting the robot's next moves.	.018	**-.792**	-.051	.004	-.060
The robot's behavior is predictable.	.067	**-.782**	.015	.019	-.019
It is easy to anticipate what will follow the robot's behavior.	.063	**-.723**	-.050	-.025	-.052
The robot's next steps are clear to me.	.136	**-.660**	-.071	.089	.114
It is difficult for me to tell what the robot will do next.	.038	**.612**	.300	.182	-.025
The robot provides cues that help predict its next actions.	.229	**-.521**	.032	.180	.214
The robot's actions are obvious.	.281	**-.521**	-.011	-.007	.067
The robot's behavior does not help predict what it will do next.	-.043	**.471**	.263	.135	-.049
The robot's overall functioning is a mystery to me.	-.008	.018	**.765**	.000	.097
It is hard to make sense of the robot's general functioning.	-.084	-.072	**.748**	.068	-.026
I am unsure what the robot does.	-.037	-.011	**.744**	-.077	-.225
It is difficult to get a clear picture of the robot's overall operations.	.012	.055	**.718**	-.117	.088
I cannot comprehend the robot's inner processes.	.032	.065	**.679**	-.040	-.013
I am confused about the robot's general objectives.	-.183	.003	**.671**	.049	.206
I cannot explain the robot's behavior.	.033	.224	**.665**	.154	-.121
It is impossible to know what the robot does.	-.124	.072	**.630**	.080	-.048
It is clear to me what the robot does.	.167	-.139	**-.519**	.260	.100
I have a clear understanding of how the robot operates in general.	.153	-.206	**-.445**	.251	-.223

Table S8. The factor pattern matrix after removing items with low factor loadings and items that loaded on multiple factors.

	I feel informed about the robot's activities.	The robot provides information about its actions.	I find the robot's explanations informative.	The robot's explanations make its actions clear to me.	I feel like the robot's explanations are useful.	The robot provides clear explanations for its actions.	The robot provides detailed explanations of its actions.	The robot's explanations for its actions are straightforward.	The robot explains complex tasks in a way that is easy to understand.	The robot conveys its overall state effectively.	The robot ensures that its users are well-informed about its activities.
I feel informed about the robot's activities.	1.000										
The robot provides information about its actions.	.683	1.000									
I find the robot's explanations informative.	.731	.740	1.000								
The robot's explanations make its actions clear to me.	.713	.759	.753	1.000							
I feel like the robot's explanations are useful.	.659	.689	.750	.739	1.000						
The robot provides clear explanations for its actions.	.663	.743	.743	.757	.680	1.000					
The robot provides detailed explanations of its actions.	.628	.704	.623	.650	.612	.622	1.000				
The robot's explanations for its actions are straightforward.	.692	.732	.738	.793	.728	.698	.641	1.000			
The robot explains complex tasks in a way that is easy to understand.	.654	.686	.705	.655	.647	.685	.651	.639	1.000		
The robot conveys its overall state effectively.	.614	.612	.644	.652	.617	.588	.511	.664	.584	1.000	
The robot ensures that its users are well-informed about its activities.	.722	.715	.718	.747	.677	.705	.640	.680	.691	.623	1.000

Table S9. Inter-Item Correlation Matrix for Factor 1 before removing highly correlated items.

ARTICLE IN PRESS

	I feel informed about the robot's activities.	I feel like the robot's explanations are useful.	The robot provides clear explanations for its actions.	The robot provides detailed explanations of its actions.	The robot's explanations for its actions are straightforward.	The robot explains complex tasks in a way that is easy to understand.	The robot conveys its overall state effectively.
I feel informed about the robot's activities.	1.000						
I feel like the robot's explanations are useful.	.659	1.000					
The robot provides clear explanations for its actions.	.663	.680	1.000				
The robot provides detailed explanations of its actions.	.628	.612	.622	1.000			
The robot's explanations for its actions are straightforward.	.692	.728	.698	.641	1.000		
The robot explains complex tasks in a way that is easy to understand.	.654	.647	.685	.651	.639	1.000	
The robot conveys its overall state effectively.	.614	.617	.588	.511	.664	.584	1.000

Table S10. Inter-Item Correlation Matrix for Factor 1 after removing highly correlated items.

	The robot's actions are obvious.	I can predict what the robot will do next.	The robot's next steps are clear to me.	I feel confident in predicting the robot's next moves.	It is easy for me to foresee the robot's future actions.	It is difficult for me to tell what the robot will do next.	The robot's behavior is predictable.	The robot provides cues that help predict its next actions.	It is easy to anticipate what will follow the robot's behavior.	The robot's behavior does not help predict what it will do next.
The robot's actions are obvious.	1.000									
I can predict what the robot will do next.	.572	1.000								
The robot's next steps are clear to me.	.617	.724	1.000							
I feel confident in predicting the robot's next moves.	.589	.740	.689	1.000						
It is easy for me to foresee the robot's future actions.	.586	.654	.639	.646	1.000					
It is difficult for me to tell what the robot will do next.	.525	.647	.614	.647	.625	1.000				
The robot's behavior is predictable.	.646	.668	.690	.671	.642	.594	1.000			
The robot provides cues that help predict its next actions.	.584	.583	.653	.571	.569	.523	.579	1.000		
It is easy to anticipate what will follow the robot's behavior.	.595	.657	.651	.646	.625	.627	.657	.542	1.000	
The robot's behavior does not help predict what it will do next.	.501	.557	.545	.542	.522	.613	.542	.479	.536	1.000

Table S11. Inter-Item Correlation Matrix for Factor 2 before removing highly correlated items.

ARTICLE IN PRESS

	The robot's actions are obvious.	The robot's next steps are clear to me.	I feel confident in predicting the robot's next moves.	It is easy for me to foresee the robot's future actions.	It is difficult for me to tell what the robot will do next.	The robot's behavior is predictable.	The robot provides cues that help predict its next actions.	It is easy to anticipate what will follow the robot's behavior.	The robot's behavior does not help predict what it will do next.
The robot's actions are obvious.	1.000								
The robot's next steps are clear to me.	.617	1.000							
I feel confident in predicting the robot's next moves.	.589	.689	1.000						
It is easy for me to foresee the robot's future actions.	.586	.639	.646	1.000					
It is difficult for me to tell what the robot will do next.	.525	.614	.647	.625	1.000				
The robot's behavior is predictable.	.646	.690	.671	.642	.594	1.000			
The robot provides cues that help predict its next actions.	.584	.653	.571	.569	.523	.579	1.000		
It is easy to anticipate what will follow the robot's behavior.	.595	.651	.646	.625	.627	.657	.542	1.000	
The robot's behavior does not help predict what it will do next.	.501	.545	.542	.522	.613	.542	.479	.536	1.000

Table S12. Inter-Item Correlation Matrix for Factor 2 after removing highly correlated items.

	I am unsure what the robot does.	It is impossible to know what the robot does.	I cannot explain the robot's behavior.	I cannot comprehend the robot's inner processes.	The robot's overall functioning is a mystery to me.	I am confused about the robot's general objectives.	It is hard to make sense of the robot's general functioning.	It is difficult to get a clear picture of the robot's overall operations.	I have a clear understanding of how the robot operates in general.	It is clear to what the robo s.
I am unsure what the robot does.	1.000									
It is impossible to know what the robot does.	.587	1.000								
I cannot explain the robot's behavior.	.638	.575	1.000							
I cannot comprehend the robot's inner processes.	.585	.520	.548	1.000						
The robot's overall functioning is a mystery to me.	.548	.614	.640	.542	1.000					
I am confused about the robot's general objectives.	.592	.598	.576	.515	.589	1.000				
It is hard to make sense of the robot's general functioning.	.583	.559	.621	.535	.557	.578	1.000			
It is difficult to get a clear picture of the robot's overall operations.	.552	.564	.577	.561	.622	.592	.542	1.000		
I have a clear understanding of how the robot operates in general.	.539	.497	.490	.508	.557	.626	.520	.570	1.000	
It is clear to me what the robot does.	.683	.597	.563	.518	.578	.586	.560	.586	.647	1.000

Table S13. Inter-Item Correlation Matrix for Factor 3, no items with high correlations were removed.

No	Item	Characteristics	
		Item Difficulty	Item Discrimination
1	The robot's overall functioning is a mystery to me.	.640	.740
2	It is hard to make sense of the robot's general functioning.	.640	.710
3	It is difficult to get a clear picture of the robot's overall operations.	.610	.730
4	I am confused about the robot's general objectives.	.660	.740
5	I am unsure what the robot does.	.660	.750
6	I cannot comprehend the robot's inner processes.	.600	.680
7	I cannot explain the robot's behavior.	.700	.740
8	It is impossible to know what the robot does.	.710	.720
9	It is clear to me what the robot does.	.660	.750
10	I have a clear understanding of how the robot operates in general.	.630	.700
11	I feel like the robot's explanations are useful.	.690	.790
12	The robot explains complex tasks in a way that is easy to understand.	.610	.770
13	The robot provides detailed explanations of its actions.	.580	.720
14	The robot provides clear explanations for its actions.	.660	.790
15	The robot's explanations for its actions are straightforward.	690	.810
16	I feel informed about the robot's activities.	.660	.780
17	The robot conveys its overall state effectively.	.690	.700
18	It is easy for me to foresee the robot's future actions.	.650	.760
19	The robot's behavior is predictable.	.690	.780
20	I feel confident in predicting the robot's next moves.	.650	.780
21	It is easy to anticipate what will follow the robot's behavior.	.660	.760
22	It is difficult for me to tell what the robot will do next.	.660	.740
23	The robot's next steps are clear to me.	.690	.800
24	The robot's actions are obvious.	.680	.720
25	The robot provides cues that help predict its next actions.	.690	.690
26	The robot's behavior does not help predict what it will do next.	.660	.660

Note: Average and range of item discrimination and item difficulty for each factor were as follows:
Factor 1: $M_{Difficulty} = .651$, $\text{range}_{Difficulty} = .600 - .710$; $M_{Discrimination} = .726$, $\text{range}_{Discrimination} = .680 - .750$.
Factor 2: $M_{Difficulty} = .654$, $\text{range}_{Difficulty} = .580 - .690$; $M_{Discrimination} = .766$, $\text{range}_{Discrimination} = .700 - .810$.
Factor 3: $M_{Difficulty} = .670$, $\text{range}_{Difficulty} = .650 - .690$; $M_{Discrimination} = .743$, $\text{range}_{Discrimination} = .660 - .800$.

Table S14. The Item Characteristics demonstrate acceptable Difficulty (.580 - .710) and strong Discrimination ($> .600$)

Constructs	Cronbach's alpha				Number of items
	Experiment 1	Experiment 2			
		English	German	Italian	
Performance Trust (MDMT)	.910	.936	.905	.872	8
Moral Trust (MDMT)	.940	.962	.921	.912	12
Acceptance (UTAUT)	.950	.963	.939	.952	25
Anxiety (UTAUT)	.730	.827	.807	.761	4
Experience with robots	.850	.876	.793	.824	8

Table S15. Reliabilities of the constructs measured in Experiment 1 and Experiment 2.

Dependent variable: Performance trust (MDMT)							
Parameter	Df	SS	MS	F value	p value	ηp^2	95 % CI
Explainability condition	1	24.30	24.35	21.26	<.001***	.06	[0.03, 1.00]
Legibility condition	1	15.50	15.48	13.51	<.001***	.04	[0.01, 1.00]
Predictability condition	1	24.60	24.62	21.49	<.001***	.06	[0.03, 1.00]
Prior experience with robots	1	19.50	19.50	17.03	<.001***	.05	[0.02, 1.00]
Explainability*Legibility	1	0.70	0.70	0.61	.437	<.01	[0.00, 1.00]
Explainability*Predictability	1	0.90	0.89	0.78	.378	<.01	[0.00, 1.00]
Legibility*Predictability	1	1.30	1.32	1.16	.283	<.01	[0.00, 1.00]
Explainability*Legibility*Predictability	2	0.00	0.00	0.00	.964	<.01	[0.00, 1.00]
Residuals	313	358.50	1.15				

Table S16. Results of the 3-way ANOVA with the manipulation of the transparency and the language of participants as independent variable, and performance trust as the dependent variable

Dependent variable: Moral trust (MDMT)							
Parameter	Df	SS	MS	F value	p value	ηp^2	95 % CI
Explainability condition	1	21.10	21.07	13.97	<.001***	.04	[0.01, 1.00]
Legibility condition	1	16.70	16.71	11.08	<.001***	.03	[0.01, 1.00]
Predictability condition	1	0.70	0.74	0.49	.485	<.01	[0.00, 1.00]
Prior experience with robots	1	38.20	38.20	25.34	<.001***	.07	[0.03, 1.00]
Explainability*Legibility	1	0.30	0.31	0.21	.651	<.01	[0.00, 1.00]
Explainability*Predictability	1	2.50	2.50	1.66	.198	.01	[0.00, 1.00]
Legibility*Predictability	1	0.00	0.01	0.01	.928	<.01	[0.00, 1.00]
Explainability*Legibility*Predictability	2	1.20	1.15	0.76	.383	<.01	[0.00, 1.00]
Residuals	313	471.90	1.51				

Table S17. Results of the 3-way ANOVA with the manipulation of explainability, legibility and predictability of the robot's behavior as independent variables, and moral trust as the dependent variable

Dependent variable: Acceptance (UTAUT)							
Parameter	Df	SS	MS	F value	P value	ηp^2	95 % CI
Explainability condition	1	17.27	17.27	19.25	<.001***	.01	[0.00, 1.00]
Legibility condition	1	4.32	4.32	4.81	.029*	<.01	[0.00, 1.00]
Predictability condition	1	5.39	5.39	6.00	.015*	<.01	[0.00, 1.00]
Prior experience with robots	1	18.27	18.27	20.35	<.001***	.01	[0.00, 1.00]
Explainability*Legibility	1	1.44	1.44	1.61	.206	<.01	[0.00, 1.00]
Explainability*Predictability	1	0.06	0.06	0.07	.795	<.01	[0.00, 1.00]
Legibility*Predictability	1	0.00	0.00	0.00	.989	<.01	[0.00, 1.00]
Explainability*Legibility*Predictability	2	0.25	0.25	0.28	.600	<.01	[0.00, 1.00]
Residuals	313	280.90	0.90				

Table S18. Results of the 3-way ANOVA with the manipulation of explainability, legibility and predictability of the robot's behavior as independent variables, and acceptance as the dependent variable

Dependent variable: Anxiety (UTAUT)							
Parameter	Df	SS	MS	F value	p value	ηp^2	95 % CI
Explainability condition	1	4.90	4.90	3.40	.066	.01	[0.00, 1.00]
Legibility condition	1	2.10	2.09	1.45	.229	<.01	[0.00, 1.00]
Predictability condition	1	0.10	0.07	0.05	.824	<.01	[0.00, 1.00]
Prior experience with robots	1	2.40	2.40	1.67	.197	.01	[0.00, 1.00]
Explainability*Legibility	1	0.70	0.68	0.474	.492	<.01	[0.00, 1.00]
Explainability*Predictability	1	0.70	0.69	0.478	.490	<.01	[0.00, 1.00]
Legibility*Predictability	1	1.10	1.12	0.77	.989	<.01	[0.00, 1.00]
Explainability*Legibility*Predictability	2	0.20	0.21	0.14	.704	<.01	[0.00, 1.00]
Residuals	313	450.40	1.44				

Table S19. Results of the 3-way ANOVA with the manipulation of explainability, legibility and predictability of the robot's behavior as independent variables, and anxiety as the dependent variable

Experiment 2

Language	Condition	No. of Participants	Age		Gender			Experience with robots	
			M	SD	Male	Female	Other	M	SD
English	Condition 1 - **High Transparency**	151	44.13	14.29	71	78	2	1.84	1.11
	Condition 2 - **Low Transparency**	149	42.70	14.57	57	90	2	1.82	1.13
German	Condition 1 - **High Transparency**	144	36.24	11.50	76	64	4	2.00	1.19
	Condition 2 - **Low Transparency**	156	36.44	12.18	86	65	5	1.98	1.11
Italian	Condition 1 - **High Transparency**	150	33.26	10.73	89	57	4	1.97	1.20
	Condition 2 - **Low Transparency**	151	32.98	9.13	73	73	5	1.98	1.23
Total		**901**	**37.62**	**12.96**	**452**	**427**	**22**	**1.93**	**1.16**

Table S20. Demographic information across the three languages and two conditions.

Standards	English	German	Italian	Acceptable	Excellent
Minimum fit function chi-square (χ^2)	1854.350	1877.748	1630.673	-	-
Degrees of freedom (df)	299	299	299		
χ^2/df	6.202	6.280	5.454	< 5.00	< 3.00
GFI	0.544	0.535	0.578	> 0.80	> 0.90
RMSEA	0.132	0.133	0.122	< 0.08	< 0.06
AGFI	0.465	0.454	0.505	> 0.80	> 0.90
NFI	0.743	0.743	0.777	> 0.85	> 0.90
CFI	0.774	0.774	0.809	> 0.90	> 0.95
TLI	0.755	0.754	0.793	> 0.90	> 0.95
IFI	0.755	0.775	0.810	> 0.90	> 0.95
SRMR	0.080	0.083	0.074	< 0.08	< 0.05

Table S21. Comparative Fit Indices of a 1-Factor model in English, German, and Italian languages in Experiment 2

Standards	English	German	Italian	Acceptable	Excellent
Minimum fit function chi-square (χ^2)	1492.204	1475.594	1298.833	-	-
Degrees of freedom (df)	298	298	298		
χ^2/df	5.007	4.952	4.359	< 5.00	< 3.00
GFI	0.623	0.626	0.658	> 0.80	> 0.90
RMSEA	0.116	0.115	0.106	< 0.08	< 0.06
AGFI	0.556	0.559	0.597	> 0.80	> 0.90
NFI	0.793	0.798	0.822	> 0.85	> 0.90
CFI	0.827	0.831	0.857	> 0.90	> 0.95
TLI	0.811	0.816	0.844	> 0.90	> 0.95
IFI	0.828	0.832	0.857	> 0.90	> 0.95
SRMR	0.076	0.078	0.066	< 0.08	< 0.05

Table S22. Comparative Fit Indices of a 2-Factor model in English, German, and Italian languages in Experiment 2

Standards	English	German	Italian	Acceptable	Excellent
Minimum fit function chi-square (χ^2)	864.821	1124.839	784.424	-	-
Degrees of freedom (df)	293	293	293		
χ^2/df	2.952	3.839	2.677	< 5.00	< 3.00
GFI	0.815	0.762	0.826	> 0.80	> 0.90
RMSEA	0.081	0.097	0.075	< 0.08	< 0.06
AGFI	0.778	0.714	0.791	> 0.80	> 0.90
NFI	0.880	0.846	0.893	> 0.85	> 0.90
CFI	0.917	0.881	0.930	> 0.90	> 0.95
TLI	0.908	0.868	0.922	> 0.90	> 0.95
IFI	0.917	0.881	0.930	> 0.90	> 0.95
SRMR	0.074	0.076	0.065	< 0.08	< 0.05

Table S23. Comparative Fit Indices of a 4-Factor model in English, German, and Italian languages in Experiment 2

Dependent variable: Performance Trust (MDMT)							
	Df	SS	MS	F value	p value	ηp^2	95 % CI
Transparency condition	1	187.90	187.89	141.46	<.001 ***	.14	[0.10, 1.00]
Language	2	190.40	95.18	71.66	<.001***	.14	[0.10, 1.00]
Transparency condition * Language	2	11.00	5.52	4.15	.016*	.01	[0.00, 1.00]
Residuals	895	1188.8	1.33				

Table S24. Results of the 2-way ANOVA with the manipulation of the transparency and the language of participants as independent variable, and Performance trust as the dependent variable

Dependent variable: Moral Trust (MDMT)							
Parameter	Df	SS	MS	F value	p value	ηp^2	95 % CI
Transparency condition	1	70.40	70.38	43.12	<.001 ***	.05	[0.03, 1.00]
Language	2	490.80	245.41	150.35	<.001***	.25	[0.21, 1.00]
Transparency condition * Language	2	22.90	11.46	7.02	<.001	.02	[0.00, 1.00]
Residuals	895	1460.80	1.63				

Table S25. Results of the 2-way ANOVA with the manipulation of the transparency and the language of participants as independent variable, and Moral trust as the dependent variable

Dependent variable: Acceptance (UTAUT)							
Parameter	Df	SS	MS	F value	p value	ηp^2	95 % CI
Transparency condition	1	59.70	59.73	43.44	<.001 ***	.05	[0.03, 1.00]
Language	2	24.40	12.18	8.86	<.001***	.02	[0.01, 1.00]
Transparency condition * Language	2	1.80	0.90	0.651	.521	<.01	[0.00, 1.00]
Residuals	895	1230.60	1.37				

Table S26. Results of the 2-way ANOVA with the manipulation of the transparency and the language of participants as independent variable, and Acceptance as the dependent variable

Dependent variable: Anxiety (UTAUT)							
	Df	SS	MS	F value	p value	ηp^2	95 % CI
Transparency condition	1	33.80	33.84	20.05	<.001 ***	.02	[0.01, 1.00]
Language	2	12.80	6.39	3.79	<.023*	.01	[0.01, 1.00]
Transparency condition * Language	2	0.30	0.13	0.08	.924	<.01	[0.00, 1.00]
Residuals	895	1510.30	1.69				

Table S27. Results of the 2-way ANOVA with the manipulation of the transparency and the language of participants as independent variable, and Anxiety as the dependent variable

Transparency Of RObots Scale (TOROS) - English Version

Georgios Angelopoulos[1,*,+], Dimitri Lacroix[2,,+], Ricarda Wullenkord[2], Alessandra Rossi[1], Silvia Rossi[1], and Friederike Eyssel[2]**

[1]Interdepartmental Center for Advances in Robotic Surgery - ICAROS, University of Naples Federico II, Naples, 80131, Italy
[2]Center for Cognitive Interaction Technology - CITEC, Bielefeld University, Bielefeld, 33619, Germany
[*]georgios.angelopoulos@unina.it
[**]dimitri.lacroix@uni-bielefeld.de
[+]these authors contributed equally to this work

Instructions

Contextualized instructions for participants:
The following statements are about the robot, its behaviors, and its functioning. Please indicate the degree to which you disagree or agree with these statements (from 1 “Strongly disagree” to 7 “Strongly agree”).

Not-contextualized instructions for participants:
Please indicate the degree to which you disagree or agree with the following statements (from 1 “Strongly disagree” to 7 “Strongly agree”).

Instructions for scoring:
By default, this questionnaire is a 7-point Likert scale (labels based on[1,2]).

1	2	3	4	5	6	7
Strongly disagree	Disagree	Somewhat disagree	Neither agree or disagree	Somewhat agree	Agree	Strongly agree

The scale can be converted as a 5-point Likert scale using the following scaling.

1	2	3	4	5
Strongly disagree	Disagree	Neither agree or disagree	Agree	Strongly agree

The authors, however, strongly recommend not to do so, as 7-point Likert scales present the best balance between ease of use, adjustment to memory span, and accuracy[1].

Instructions for administration:
The order of the presented items should be ideally randomized.

Instructions for scoring:
Subscale (dimension) scores are calculated by averaging the ratings of the items of each subscale. A composite score of transparency can be calculated with the average of the three subscales.

The items:

Factors	Items
Illegibility	The robot's overall functioning is a mystery to me. It is hard to make sense of the robot's general functioning. It is difficult to get a clear picture of the robot's overall operations I am confused about the robot's general objectives. I am unsure what the robot does. I cannot comprehend the robot's inner processes. I cannot explain the robot's behavior. It is impossible to know what the robot does. It is clear to me what the robot does. (R) I have a clear understanding of how the robot operates in general. (R)
Explainability	I feel like the robot's explanations are useful. The robot explains complex tasks in a way that is easy to understand. The robot provides detailed explanations of its actions. The robot provides clear explanations for its actions. The robot's explanations for its actions are straightforward. I feel informed about the robot's activities. The robot conveys its overall state effectively.
Predictability	It is easy for me to foresee the robot's future actions. The robot's behavior is predictable. I feel confident in predicting the robot's next moves. It is easy to anticipate what will follow the robot's behavior. It is difficult for me to tell what the robot will do next. (R) The robot's next steps are clear to me. The robot's actions are obvious. The robot provides cues that help predict its next actions. The robot's behavior does not help predict what it will do next. (R)

Note: (R) indicates reverse-coded items that require score inversion before analysis.

References

1. Taherdoost, H. What is the best response scale for survey and questionnaire design; review of different lengths of rating scale/attitude scale/likert scale. *Hamed Taherdoost* 1–10 (2019).
2. Wade, M. V. *et al.* Likert-type scale response anchors. *Clemson international institute for tourism & research development, department parks, recreation tourism management. Clemson Univ.* 4–5 (2006).

Transparency Of RObots Scale (TOROS) - German Version

Georgios Angelopoulos[1,*,+], Dimitri Lacroix[2,,+], Ricarda Wullenkord[2], Alessandra Rossi[1], Silvia Rossi[1], and Friederike Eyssel[2]**

[1]Interdepartmental Center for Advances in Robotic Surgery - ICAROS, University of Naples Federico II, Naples, 80131, Italy
[2]Center for Cognitive Interaction Technology - CITEC, Bielefeld University, Bielefeld, 33619, Germany
[*]georgios.angelopoulos@unina.it
[**]dimitri.lacroix@uni-bielefeld.de
[+]these authors contributed equally to this work

Instruktionen

Kontextabhängig:
Die folgenden Aussagen beziehen sich auf den Roboter, seine Verhaltensweisen und seine Funktionsweise. Bitte geben Sie an, inwieweit Sie diesen Aussagen zustimmen oder sie ablehnen (von 1 “lehne stark ab” bis 7 “stimme stark zu”).

Nicht kontextualisiert:
Sie diesen Aussagen zustimmen oder sie ablehnen (von 1 “lehne stark ab” bis 7 “stimme stark zu”).

Hinweis:
Standardmäßig handelt es sich bei diesem Fragebogen um eine 7-Punkte-Likert-Skala[1,2]).

1	2	3	4	5	6	7
lehne stark ab	lehne ab	lehne etwas ab	stimme weder zu noch lehne es ab	stimme eher zu	stimme zu	stimme stark zu

Die Skala kann anhand der folgenden Skalierung in eine 5-Punkte-Likert-Skala umgewandelt werden.

1	2	3	4	5
lehne stark ab	lehne ab	stimme weder zu noch lehne es ab	stimme zu	stimme stark zu

Die Autoren empfehlen jedoch dringend, dies nicht zu tun, da 7-Punkte-Likert-Skalen die beste Balance zwischen Benutzerfreundlichkeit, Anpassung an die Gedächtnisspanne und Genauigkeit bieten[1].

Anweisungen für die Durchführung:
Die Reihenfolge der präsentierten Items sollte idealerweise randomisiert sein.

Hinweise zur Auswertung:
Die Subskalen (Dimensionen) werden durch Mittelung der Bewertungen der Items der einzelnen Subskalen berechnet. Aus dem Durchschnitt der drei Subskalen kann ein zusammengesetzter Wert für die Transparenz berechnet werden.

Die Items:

Faktoren	Die Items
Unschärfe	Die allgemeine Funktionsweise des Roboters ist für mich ein Rätsel. Es ist schwierig, die allgemeine Funktionsweise des Roboters zu verstehen. Es ist schwierig, sich ein klares Bild von der allgemeinen Funktionsweise des Roboters zu machen. Ich bin über die allgemeinen Ziele des Roboters verwirrt. Ich bin unsicher, was der Roboter macht. Ich kann die inneren Vorgänge des Roboters nicht nachvollziehen. Ich kann mir das Verhalten des Roboters nicht erklären. Es ist unmöglich zu wissen, was der Roboter tut. Es ist mir klar, was der Roboter macht. (R) Ich habe eine klare Vorstellung davon, wie der Roboter im Allgemeinen funktioniert. (R)
Erklärbarkeit	Ich habe das Gefühl, dass die Erklärungen des Roboters nützlich sind. Der Roboter erklärt komplexe Aufgaben auf eine leicht verständliche Weise. Der Roboter gibt detaillierte Erklärungen für seine Handlungen ab. Der Roboter liefert klare Erklärungen für seine Handlungen. Die Erklärungen des Roboters für seine Handlungen sind einfach. Ich fühle mich über die Aktivitäten des Roboters informiert. Der Roboter vermittelt seinen allgemeinen Zustand effektiv.
Vorhersehbarkeit	Es fällt mir leicht, die zukünftigen Aktionen des Roboters vorherzusehen. Das Verhalten des Roboters ist vorhersehbar. Ich bin zuversichtlich, dass ich die nächsten Schritte des Roboters vorhersagen kann. Es ist leicht vorauszusehen, was auf das Verhalten des Roboters folgen wird. Es ist schwierig für mich zu sagen, was der Roboter als nächstes tun wird. (R) Die nächsten Schritte des Roboters sind für mich klar. Die Aktionen des Roboters sind offensichtlich. Der Roboter gibt Hinweise, die helfen, seine nächsten Handlungen vorherzusagen. Das Verhalten des Roboters hilft nicht dabei vorherzusagen, was er als nächstes tun wird. (R)

Hinweis: (R) kennzeichnet umgekehrt codierte Items, deren Bewertungen vor der Analyse invertiert werden müssen.

References

1. Taherdoost, H. What Is the Best Response Scale for Survey and Questionnaire Design; Review of Different Lengths of Rating Scale / Attitude Scale / Likert Scale. *Int. J. Acad. Res. Manag.* **8**, 1–10 (2022).
2. Wade, M. V. *et al.* Likert-type scale response anchors. *Clemson international institute for tourism & research development, department parks, recreation tourism management. Clemson Univ.* 4–5 (2006).

Acknowledgements

Die Autoren danken Lena Schubert für ihren Beitrag zur Übersetzung der Skala in die deutsche Sprache.

Transparency Of RObots Scale (TOROS) - Italian Version

Georgios Angelopoulos[1,*,+], Dimitri Lacroix[2,,+], Ricarda Wullenkord[2], Alessandra Rossi[1], Silvia Rossi[1], and Friederike Eyssel[2]**

[1]Interdepartmental Center for Advances in Robotic Surgery - ICAROS, University of Naples Federico II, Naples, 80131, Italy
[2]Center for Cognitive Interaction Technology - CITEC, Bielefeld University, Bielefeld, 33619, Germany
[*]georgios.angelopoulos@unina.it
[**]dimitri.lacroix@uni-bielefeld.de
[+]these authors contributed equally to this work

Istruzioni

Istruzioni contestualizzate per i partecipanti:
Le seguenti affermazioni riguardano i robot, i loro comportamenti e le loro funzionalità. Per piacere indicate il grado con cui siete in disaccordo o in accordo con queste affermazioni (da 1 “Fortemente in disaccordo” a 7 “Fortemente d’accordo”).

Istruzioni non contestualizzate per i partecipanti:
Per piacere indicate il grado con cui siete in disaccordo o in accordo con queste affermazioni (da 1 “Fortemente in disaccordo” a 7 “Fortemente d’accordo”).

Istruzioni per la valutazione:
i base questo questionario è una scala Likert a 7 punte (le etichette sono basate su[1,2]).

1	2	3	4	5	6	7
Fortemente in disaccordo	In disaccordo	Abbastanza in disaccordo	Né d’accordo né in disaccordo	Abbastanza d’accordo	D’accordo	Fortemente d’accordo

Tuttavia la scala qui proposta (scala Likert a 7 punte) può essere convertita in nella seguente versione ridotta (scala Likert a 5 punte).

1	2	3	4	5
Fortemente in disaccordo	In disaccordo	Né d’accordo né in disaccordo	D’accordo	Fortemente d’accordo

Gli autori raccomandano l’utilizzo della versione di scale Likert a 7 punte in quanto rappresentano il miglior bilanciamento tra facilità d’uso, adeguamento alla capacità di associazione, e accuratezza[1].

Istruzioni per la somministrazione:
Gli autori suggeriscono di presentare gli elementi della scala in ordine casuale.

Istruzioni per la valutazione:
I punteggi della Sottoscala (dimensione) sono calcolati facendo una media semplice della valutazione dei singoli elementi di ogni sottoscala. Il punteggio composto della trasparenza può essere calcolato come la media delle tre sottoscale.

Le voci:

Fattori	Le voci
Illeggibilità	Il funzionamento generale del robot è un mistero per me. E' difficile capire il generale funzionamento del robot. E' difficile avere una chiara visione delle operazioni generali del robot. Sono confuso sul obiettivo generale del robot. Non sono sicuro di cosa faccia il robot. Non capisco quali siano i processi interni del robot. Non so spiegare il comportamento del robot. E' impossibile sapere cosa il robot faccia. Mi è chiaro cosa il robot faccia. Ho una chiara comprensione sul come il robot operi in generale. (R) Ho l'impressione che le spiegazioni del robot siano utili. (R)
Spiegabilità	Il robot spiega task complessi in un modo semplice da capire. Il robot da spiegazioni dettagliate delle sue azioni. Il robot da chiare spiegazioni delle sue azioni. Le spiegazioni date dal robot sulle sue azioni sono dirette. Mi sento informato riguardo le attività del robot. Il robot comunica il suo stato generale in maniera effettiva. E' facile per me prevedere le future azioni del robot.
Prevedibilità	Il comportamento del robot è prevedibile. Mi sento sicuro nel predire i movimenti successivi del robot. E' facile anticipare cosa avverrà dal comportamento del robot. E' difficile per me dire cosa il robot farà successivamente. (R) Le prossime azioni del robot sono chiare per me. Le azioni del robot sono scontate. Il robot da indizi che aiutano a predire le sue successive azioni. Il comportamento del robot non aiuta a predire cosa farà successivamente. (R)

Nota: (R) indica voci con codifica inversa, i cui punteggi devono essere invertiti prima dell'analisi.

References

1. Taherdoost, H. What Is the Best Response Scale for Survey and Questionnaire Design; Review of Different Lengths of Rating Scale / Attitude Scale / Likert Scale. *Int. J. Acad. Res. Manag.* **8**, 1–10 (2022).

2. Wade, M. V. *et al.* Likert-type scale response anchors. *Clemson international institute for tourism & research development, department parks, recreation tourism management. Clemson Univ.* 4–5 (2006).

Acknowledgements

Gli autori ringraziano Francesco Vigni per il suo contributo alla traduzione della scala in lingua Italiana.