Comparing a mentalist and an interactionist approach for trust analysis in Human-Robot Interaction

Marc Hulcelle
marc.hulcelle@telecom-paris.fr
Télécom-Paris, Institut Polytechnique de Paris
Palaiseau, France

Giovanna Varni
giovanna.varni@unitn.it
Department of Information Engineering and Computer Science (DISI), University of Trento
Povo, Italy

Nicolas Rollet
nicolas.rollet@telecom-paris.fr
Télécom-Paris, Institut Polytechnique de Paris
Palaiseau, France

Chloé Clavel
chloe.clavel@telecom-paris.fr
Télécom-Paris, Institut Polytechnique de Paris
Palaiseau, France

ABSTRACT

Trust is an important aspect of a human-robot interaction (HRI) as it mitigates the performance of many activities. Users' trust may be impacted when robots make mistakes. To be able to properly time trust-reparation actions, robots should detect trust variations during the interaction. There are very few computational models of trust for such a task. The existing ones relied on either Psychological or Sociological theories that gave place to different definitions and analysis tools. We can distinguish two main approaches in the trust literature: the mentalist and the interactionist one. In this paper, we compare both approaches for trust detection, and explore how the adoption of two different assessment tools on an HRI dataset may lead to different results. We identify criteria that set them apart, and provide guidelines on the possibilities that each approach offers depending on the target computational model of trust.

CCS CONCEPTS

• **Computer systems organization → Robotics**.

KEYWORDS

HRI, Trust, Interactional Sociology, Psychology, Methodologies

ACM Reference Format:
Marc Hulcelle, Giovanna Varni, Nicolas Rollet, and Chloé Clavel. 2023. Comparing a mentalist and an interactionist approach for trust analysis in Human-Robot Interaction. In *Proceedings of 11th International Conference on Human-Agent Interaction (HAI '23).* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Trust, as a fundamental concept of the human-robot interaction, affects the acceptance of the robot and the interaction task performance [1, 18, 21]. Studies on trust in HRI have historically been grounded in Psychology, more specifically with a *mentalist* approach that defines trust as a state of mind of the user. Psychological studies focus on factors - robot, user, and environment-related - that affect trust, and show that robot-related factors, such as its design and behavior, have a greater impact [18, 19, 26, 41]. Some other studies focus on the users' display of trust to monitor trust externally (e.g. [22]). Among those, a few of them adopt an *interactionist* approach through Interactional Sociology [3, 15]. While the object under study is the same, that is whether participants trust the robot or not, the methodology differs. In these studies, trust is defined as a process and is made visible by the users through their behaviors according to normative expectations [11, 12, 14].

During interactions, failures are bound to happen. Trust lowered by these failures should be repaired, as a risk reassurance between a trustor and a trustee, by adapted and appropriately timed reparation strategies [7, 27]. To achieve this, the robot should be able to detect variations of trust at a fine-grained level during the interaction by relying on the analysis of the users' behaviors. There are very few studies that proposed computational models of trust for this task [22, 42]. Such analysis relies on a few constraints to build the ground truth. First, the analysis should be performed through an automatic segmentation, generally done through fixed-length sliding windows whose length should be determined. Second, each segment should be annotated to build the ground truth for the model to be trained on. Given the regularity of the trust assessment, annotations would ideally be collected by an external observer so as not to disrupt the flow of the interaction.

In this paper, we carry out a comparative analysis - first through a study of their theoretical frameworks, then through an experimental analysis of annotations - of both the interactional sociological perspective, which we refer to as the interactionist approach in the following, and the psychological perspective that adopts a mentalist approach, for a fine-grained analysis of trust during the interaction. We investigate the following research questions:

(1) **RQ1**: Can we identify criteria to differentiate both approaches based on their theoretical framework and trust assessment tools' methodologies ?
(2) **RQ2**: According to the objective of a study, can we identify the possibilities that each approach offers to build a computational model of trust ?

HAI '23, December 4–7, 2023, Gothenburg, SE

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

We conduct an analysis of annotation differences of two assessment tools from both approaches on an HRI dataset. To be able to compare them during our study, we will discuss the few adaptations that we made to fit our task of a fine-grained analysis of trust.

2 A MENTALIST APPROACH

2.1 Theoretical framework

Grounding on Psychological theories, trust in HRI is described as a mental state, in which users have a certain set of expectations - either positive or negative - towards the robot's technical and social skills. One of the most commonly used definition in HRI characterized trust as *"psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another"* [32]. Most definitions, such as the previous one, rely on the idea that users form a mental model of the robot's capacity to handle uncertain situations, and act benevolently [19, 23, 36]. Users thus expect the robot to not harm their physical/mental well-being, respect each other's interests, and deal appropriately with uncertain situations. Building on this background, two components of trust were distinguished: *cognitive trust* (CT) - the *"self-efficacy to rely on capabilities and reliabilities of a specific party"* - and *affective trust* (AT) - the *"self-efficacy on the party based on human affective responses to the behavior of the party"* [36]. Trust is, therefore, a combination of the user's mental projection of the robot's capacities and an affective response to these. Psychological definitions also show that trust is affected by the observation of present events and the user's projection of future events. This link between trust and the uncertainty generated by the projection of future events is highlighted in [3] that defines trust as *"a process of uncertainty reduction, the ultimate goal of which is to reinforce assumptions about a partner's dependability with actual evidence from the partner's behavior"*. Interestingly, while the last definition mentions the user's mental model of their partner's dependability, it also mentions the reliance on perceptible behaviors and behavioral proofs of this dependability, which leads us to the interactionist approach that we later describe in Section 3.

2.2 Mentalist trust assessment tools

Available trust assessment tools in HRI built on a psychological background require participants to answer questions about their mental representation of the robot they will interact or have interacted with. The most used questionnaires are the Interpersonal Trust Scale (ITS) [31], the Negative Attitudes towards Robots Scale (NARS) [40], Godspeed [2], and the Robot Trust Scale (RTS) [36]. While NARS can be useful to assess a participant's negative preconception of the robot to interact with, NARS and ITS questionnaires are highly correlated with pre-interaction trust measures, but not post-interaction trust measures [36]. The RTS can be filled by participants before and after an interaction with a robot to assess their trust. This scale focuses on antecedents and measurable factors of trust related to the human, robot, and environment. The scale comprises 40 items, but a smaller subset of 14 items can be used for a faster assessment [36]. Each item represents the participant's expectation of the robot's behavior given their mental model of the robot - e.g. *"What % of the time will this robot act consistently"*. Answers are given in the form of an 11-point Likert scale, from 0 to 100. The final trust score is the average of all individual items' score. The scale encompasses items that relate either to the robot's perceived technical or social skills. Some items can be interpreted in both ways. For instance, "acting consistently" can either relate to the predictability of the output of the task the robot is working on - e.g. baking cookies - or to the consistency of its displayed personality - e.g. being friendly, then becoming overly sarcastic would be inconsistent.

3 AN INTERACTIONIST APPROACH

3.1 Theoretical framework

Interactional Sociology relies on the observability of trust within the interaction, which is made visible by the participants themselves through their behaviors [11, 12, 14]. Trust is thus a result of the state of the interaction, and is oriented towards both the content and the format of the interaction. In a trusting state, participants tend to behave in a way so that the interaction is fluid and proceeds towards its objective [8]. It is observable on different bases: e.g. trust in the robot's capacity to maintain a fluid and progressive interaction, in its knowledge, its skill in accomplishing a specific action at a given moment. Interactional trust was defined as a *"form of affiliation and credit characterized by a set of behaviors that are intentional or not, expressive or propositional"* [15]. This definition relies on concepts of *alignment* [39] - *i.e.* complying with the *trajectory* (sequential progression) of the interaction -, *affiliation* - claiming access to and understanding the partner's stance, and endorsing their perspective - [39], and *credit* [6, 28] given to the robot's *competence* [10]. Credit is the recognition of the relevance and suitability of the partner's message or social behavior in the interaction's context.

3.2 Interactionist trust assessment tools

To the best of our knowledge, the coding scheme "Trust in hUman Robot INteraction" (TURIN) [15] is the only trust assessment tool grounded on Interactional Sociology that exists in HRI. This annotation scheme stems from methodologies offered by ethnomethodology and interactional sociology [9, 11, 13, 14], in particular conversation analysis. Trust is thus analyzed in a continuous way throughout the entire interaction. Annotators rely on salient behaviors made observable by the participants themselves to assign a trust label. Annotators are instructed to annotate social behaviors related to trust. The coding scheme enables the unitizing of the interaction into segments of coherent trust categories, and provides three labels: *"Trusting"*, *"Neutral"*, and *"Mistrusting"*. The segmentation is first carried out on behavioral units that indicate behavior changes based on the ongoing interactional process. Consecutive units that share the same trust category are aggregated together to form a single coherent segment. Then, more detailed annotations are given to describe which behaviors (e.g. gaze, intonation) are indicative of the trust label with labels from the "Social Interaction Form" subcategory. Further subcategories "Interaction Content" (e.g. alignment), "Benevolence" (e.g. self-disclosure), and "Integrity" (e.g. honesty) bring additional hints on how participants exhibit trust. These subcategories allow to annotate observable behaviors and social phenomenons (e.g. alignment, compliance). According to definitions given in Section 3.1, "Trusting" labels are assigned to

segments in which participants exhibit behaviors that show alignment, affiliation, accept vulnerability or validate the robot's skills. "Mistrusting" labels are given to segments in which participants display the opposite type of behavior: disalignment, disaffiliation, or doubting the robot's skills [39]. The "Neutral" label is assigned to all other segments of the interaction. In that sense, it is an absence of salient behavior that the annotators observed.

4 A PILOT STUDY

In this section, we compare the RTS and TURIN through a pilot study to investigate which approach is best suited to build machine-learning models for a fine-grained analysis of trust throughout the interaction, and investigate how both approaches can complement each other.

Dataset. We chose the Vernissage corpus as a test-bed for our study [17]. Among all publicly available HRI datasets, Vernissage is the only one adopting a semi-structured experimental scenario in which participants are unconstrained in their behavior. Vernissage is composed of 10 group interactions involving two human participants and a NAO robot. Participants follow three predefined experimental phases: the robot first presents five paintings behind it (*vernissage*), then asks the participants to present themselves with more information than simply their name (*self-presentation*), and finally quizzes them on art (*art-quiz*).

Selected tools. We chose the RTS as an annotation tool for the mentalist approach. The RTS is the most comprehensive among other previously cited tools [36] and is the only one to be correlated with post-trust interaction [37], showing that it is able to measure trust variations. We used the RTS reduced to 14 items version. We chose the reduced version of the RTS to limit the cognitive load of the annotation process. Furthermore, many items from the full scale focus on robots' technical and social skills that are too general - e.g. "Protecting people", "Warning people of potential risks in the environment", "Performing many functions at one time". As such, they are irrelevant in the context of the Vernissage experimental scenario. Among the 14 items, we considered the items "perform exactly as instructed" and "follow directions" to be unrelated to the task since the robot acts as an art guide, and is thus mostly in charge of the conversation and never has to follow any of the participant's instructions.

For the interactionist approach, we chose TURIN since it's the only publicly available tool from this approach to the best of our knowledge.

Adapting the tools. We operate changes on the RTS for our task given its constraints. One of the constraints is that annotations are required to be collected during regular and small time-frames. The RTS is generally used at the beginning and end of the interaction as it takes time to fill given the amount of items. Interrupting the interaction in such a way would disrupt its flow. As a consequence, there are no publicly available dataset that includes annotations collected in such way. Annotations should be conducted from an external observer's point of view. We thus had to adjust the RTS' point-of-view since it was not designed to assess participants' trust by an external observer. As there is no mentalist assessment tool with a third-person view and for the tool to fit our task, we asked the annotator to consider them-self as a bystander of the interaction. From the observation of the participants' behaviors and reactions to the robot, each expert built their own perception of the robot which they used to fill the RTS. Hung et al., for instance, performed such translation of questionnaires in a third-person point of view to study the cohesion of small human groups based on nonverbal audio-visual behaviors [16]. We did not ask the annotator to try to infer participants' state of mind from their behavior as the RTS items relate to perceptions of the robot's skills and do not relate to user-centered criteria. Considering this issue, and to avoid interpretation bias, we asked the expert to annotate its own perception of trust towards the robot.

For the annotation comparison study, we adjusted the TURIN's unitizing method to fit the task's constraints. First, we changed TURIN's unitizing method by collecting annotations based on fixed-length windows, even though TURIN specifies a unitizing method that relies on the aggregation of moments of coherent trust category. Even though this study's unitizing method is not grounded in the undergoing interactional processes, these processes are still visible within a segment. Thus, we decided here to focus on the trust category that is dominant within a segment, and TURIN's sub-categories that are made visible by users within a segment. The annotation length from TURIN's subcategories does not necessarily match the segment's length in the original approach. Subcategory items are used originally to describe behaviors that happen during a time-frame that is relevant in the interaction's structure. The annotator has to choose at most 2 annotations from the "Social Interaction Form" sub-category, at most 2 annotations from the "Interaction Content" category, at most 1 from the "Benevolence" one, and at most one from the "Integrity" one. We limited the annotations in such a way to only report behaviors that were the most salient inside units and avoid having too many annotations about punctual behaviors.

Procedure. Five experts[1] in HRI - that we name A, B, C, D, and E - participated in the study. We collected annotations on the first three minutes of the 10 interactions, focusing on the first three paintings of the vernissage phase for this preliminary study. Expert A annotated all 10 interactions with both approaches. They had previous training and experience with TURIN before conducting our study's task. Annotations were collected from expert A with a time lapse of a week between both methods to reduce the influence of one task on the other. Experts B and C annotated five different interactions with no overlap between them with the interactionist tool. Experts D and E did the same with the mentalist one. The experts annotated fixed-length windows of 10 seconds, yielding a total of 18 segments per interaction. When choosing the windows' length, we reached a compromise close to a few speaking turns for TURIN to still be able to highlight behaviors in a relevant time-frame, and long enough for the annotator's representation of RTS items to evolve. While the RTS can be filled by annotators right after being presented and items' definition clarified, TURIN requires annotators to be trained before using it. Experts were trained during a two hours-long annotation workshop on videos outside of the pool of their assigned interactions

[1]Annotators had different expertise level in robotics, with backgrounds in affective computing, and machine learning.

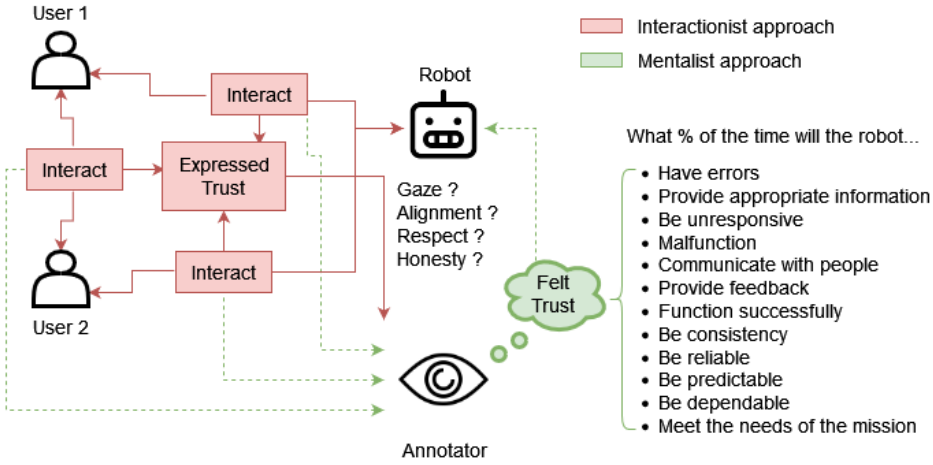

Figure 1: Through interactions between all group members, the group makes observable to the annotator the expressed group trust. In the interactionist approach, the annotator relies on observable and tangible evidence of members' behaviors to assign a trust label from TURIN. In the mentalist approach, the annotator relies on its interpretation of these interactions and its own perception to assign a score to each criteria of the RTS.
Full lines: explicit behavior. Dashed lines: perceptions.

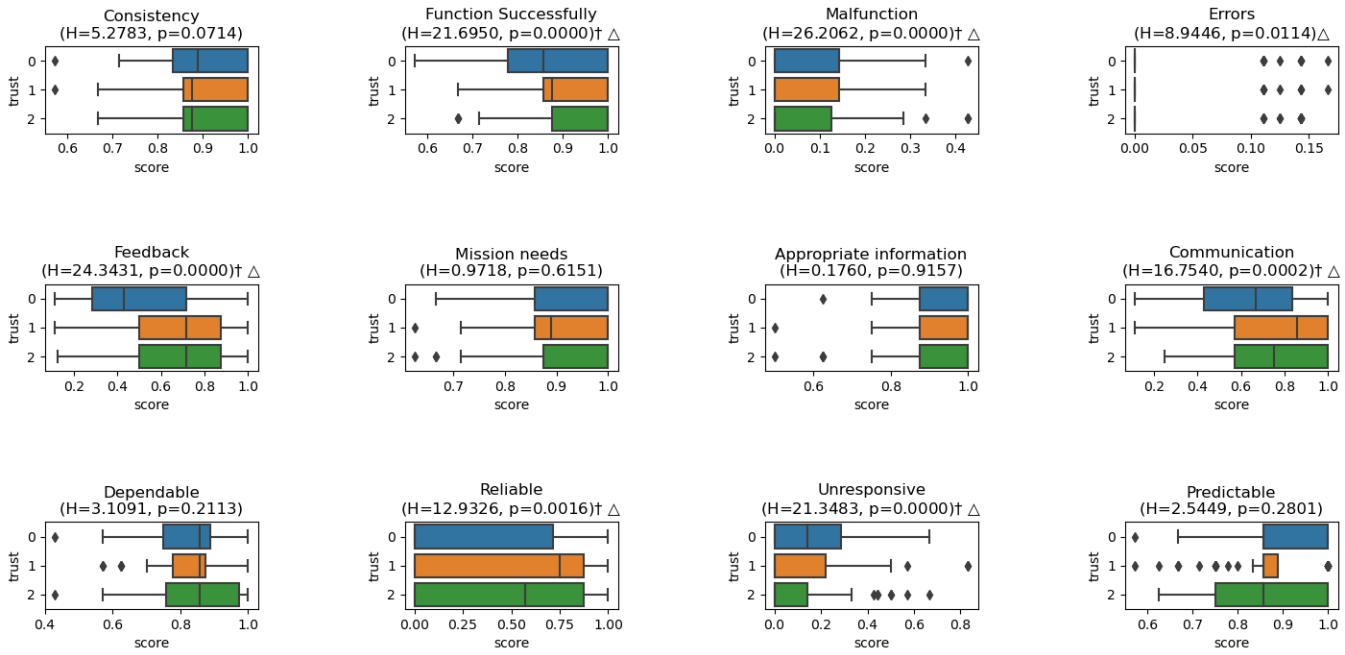

Figure 2: Score distribution of the 12 items for the RTS and correlation with TURIN trust annotations.
Trust values: 0="Mistrusting", 1="Neutral", 2="Trusting".
† : significant score distribution difference between Mistrusting and Neutral segments.
△ : significant score distribution difference between Mistrusting and Trusting segments.
○ : significant score distribution difference between Neutral and Trusting segments.

We applied all previously described adaptations of the tools as a way for them to meet a common ground for comparison purposes. As there are no publicly available HRI datasets that contains RTS annotations, we had to make more adaptations to it than TURIN for this pilot study. Figure 1 provides a summary of the annotation process for the interactionist and the mentalist approach.

4.1 Results

4.1.1 Comparing approaches. To compare both approaches, we aggregated segments based on their assigned TURIN trust label. To aggregate the annotations from all experts, we first rescale the scores of each item from the RTS for each annotator (A, B, and C). We perform a Shapiro-Wilk test to assess whether each item's score for each annotator come from a normal distribution [38]. The results show that none come from a normal distribution, $p < .001$. We thus operate a min-max scaling of each of the RTS items' score for each annotator. Then, depending on the assigned TURIN trust label of the segment, we searched for statistical differences in the mean score of each of the RTS items . We first applied a Kruskal-Wallis test [20]. If the test reveals significant difference, it is followed by a post-hoc Dunn test with Bonferroni correction [5]. We plot the score distributions and report all results of the statistical tests in Figure 2.

First, we observe statistically significant differences between "Mistrusting" and "Trusting" segments for items "Function successfully", "Malfunction", "Errors", "Feedback", "Communication", "Reliable", and "Unresponsive", $p < .05$. This is due to the fact that participants strongly react to faulty behaviors from the robot, for instance when the robot ignores the answer of participants after asking them a question, or when it fails to recognize the participants' name. Except for item "Errors", the test reveals significant differences between "Mistrusting" and "Neutral" segments for all previously cited items.

Other items "Consistency", "Mission needs", "Appropriate information", "Dependable", and "Predictable" appear independent of TURIN trust labels: these items can take any value, TURIN labels will not necessarily reflect the RTS trust label. Looking at annotations closely, some participants still align and comply with the robot even when it displays faulty behavior, while others might still express doubt towards the robot even if it functions perfectly well.

4.1.2 Comparing annotators. We then studied the inter-rater agreement (IRA) between annotator A and annotators B, C, D, and E. For TURIN, we computed the Cohen's Kappa [4] between A and B for each interaction and the overall agreement on all interactions, and did similarly for A and C since there is no overlap between B and C. The overall Cohen's Kappa is rather weak with TURIN, 0.37 between A and B, 0.35 between A and C. This can be explained by the choice of using fixed-length windows to collect annotations and compare both approaches. Since windows are not rooted in the structure of the interaction, annotators may decide to highlight different phenomenon that may happen within these. When investigating the mismatches between annotators, we observe that the highest number of mismatches appear when one of the annotators chose the "Neutral" label. Assigning the "Neutral" label requires the annotators to evaluate whether behaviors are not significant enough to be considered signs of trust or mistrust. This highlights the difficulty of the annotation task given the constraint on windows' length.

We then computed the Cohen's kappa for TURIN's subcategories "Interaction Form" and "Interaction Content". For "Interaction Form", the Cohen's Kappa is poor, $\kappa = 0.13$ between A and B, $\kappa = 0.09$ between A and C. The low value can be explained by the constraints of the task: the experts had to choose at most two from many items that were dominant in the segment. They reported during a post-annotation interview that this limitation in the choice of the item made the item selection difficult given the length of the segment. They reported that they often would have like adding a third item. The experts mostly annotated with items "Gaze", "Facial expression", and "Intonation" For "Interaction Content", the Cohen's Kapp is poor between A and B, $\kappa = 0.08$, but weak between A and C, $\kappa = 0.24$. The low Kappa value can be explained by the fact that this subcategory is a descriptor of the "Interaction Form" subcategory. As such, it relies on previous item selection during this task and depends on what the annotator chose to focus on in the segment. Items that were the most used for this subcategory were "Alignment", "Approval" and "Compliance"

For the RTS, we considered each category's set of values through the interaction as time-series. Therefore, we computed the Cramer's V correlations between annotators for each category, and averaged them for an interaction as we observed no negative correlations between annotators. Correlations are pretty low, $V = 0.16$ between A and D, and $V = 0.22$ between A and E. This result highlights the subjectivity of the task, as expected from a mentalist approach. Categories "Errors", "Mission Needs", "Reliable", and "Dependable" yield the lowest correlations. This is explained by the content of the interaction: the robot acts as an art guide, and as such, the mission needs may not appear very clear to annotators. As there is no explicit vulnerability in the experimental scenario of the dataset, dependency and reliance on the robot from participants may also appear unclear. Category "Errors" low correlations might be explained by different annotation practices from annotators : expert E used higher ranges of scores than expert A, thus increasing the size of the contingency table.

During post-annotation interviews, all annotators pointed out that interactions 2, 3, 9, and 10 were significantly harder to annotate than the others. Annotators reported that the discrepancy between the enthusiasm of some participants and the faulty behaviors of the robot made the annotation task difficult for these interaction. Annotators were also unsure whether some participants were sometimes acting sarcastically or not in these interactions. Moreover, in interaction 9, one of the participants has trouble understanding Nao and often asks the other participant to translate in their native language. While they point out Nao's faulty behaviors, they still show signs of trust by asking it to slow down for instance. Because of this, annotators expressed difficulty in choosing the appropriate TURIN trust label.

5 DISCUSSION

To answer the **RQ1**, we identified 4 criteria, which we detail in the following sections, on which both approaches differ from our

	Time-framing			Orientation		Generalization		Scalability	
	BU	ST	EI	Data-driven	Theoretical-framework-driven	Specific	Generic	Individual	Group
Mentalist			X		X		X	X	X
Interactionist	X	X		X		X		X	X

Table 1: Summary of the comparison of the mentalist and interacionist approach based on 4 criteria. BU: Behavioral Unity. ST: Speaking Turns. EI: Entire interaction.

theoretical and annotation comparison study : *orientation*, *generalization capability*, *time-framing*, and *scalability*. A summary can be found in Table 1.

5.1 Differentiating criteria

5.1.1 Orientation. First, the approaches' *orientation* diverges, which corresponds to the preference for theoretical-framework-driven or data-driven tools. Our theoretical analysis showed that the mentalist approach has led to the creation of rather **theoretical-framework-driven** assessment tools, while the interactionist assessment tools are more **data-driven** as study results solely rely on the close examination of users' behaviors that emerge from the data. There is a tension between considering trust as a mental state, and admitting that given that the participants do not have access to their partner's brain, it is towards the observability of this supposed state that the participants are oriented. They can nonetheless try to infer it through the partner's behaviors and decisions [22]. However, past studies show that even when a robot displays faulty behaviors that are detrimental to trust, participants sometimes still decide to follow the robot's advice, e.g. during a fire alarm [30, 35]. This is confirmed by our study. Indeed, in the Vernissage scenario, the robot often shows difficulties in grasping the users' names and asks the participants to repeat. RTS annotations indicate that participants' trust is low after that. But TURIN annotations show that participants still trust the robot to refer to the correct paintings right after the mistake. This discrepancy between the user's expected behavior and its actual behavior shows the difficulty of inferring the user's mental model of the robot's trustworthiness [19, 37].

5.1.2 Generalization capability. Next, the approaches differ on their *generalization capability*, based on how specific or generic their analysis is according to the interaction's task. The mentalist approach is **quite generic** and not dependent on the interaction history. Assessment tools' items cover a wide range of concepts relating to trust that do not depend on the interaction's task. However, the pilot study shows that some items from the RTS are very similar, such as "performing exactly as instructed" and "following directions". Given this and their potential double interpretation, the study of the robot's behavioral factors affecting users' trust can be difficult depending on the interaction's confounders. As for the interactionist approach, the small time-framing makes the interactional history important during the analysis, making this approach **context-sensitive and non-generic**. Depending on which interactional process is being studied, and therefore the time-frame of analysis, the interpretation can lead to different labels assigned by the annotator. Some behaviors can also be interpreted in both ways. For instance, after the robot fails to first understand the name of a participant, the participant may repeat itself. By doing so, the participant highlights the robot's failure and disrupts the interaction's fluidity. But, by repeating, they start an error reparation process and thus reveal that they still trust the robot to understand their name.

5.1.3 Time-framing. The approaches' *time-framing*, the optimal time-interval necessary for the analysis, also differs. Our theoretical analysis showed that the interactionist approach's time-framing is **close to one or a short series of behavioral units**, since it heavily relies on the interactional history. This approach is highly dependent on human unitizing and requires more training before being used. In our study, a post-annotation interview revealed that using unitizing that is not grounded on the interaction dynamics can lead to difficulties on the choice of the trust label. On the other hand, the mentalist approach has a much longer time-framing. Indeed, as trust assessment tools are questionnaires that are time-consuming to fill, measures are generally conducted at the end and beginning of the interaction. Their time-framing is thus generally **the entire interaction** so that all criteria have enough time to evolve, and **sometimes several interactions** depending on the criterion. This reduces the possibilities to investigate the evolution of trust within the interaction. The focus is on the participants' representations of the robot's *global* capacities and not the participants' behaviors.

5.1.4 Scalability. Last, the approaches diverge on their *scalability*, the ability to be used for the analysis of a single user or larger groups. The theoretical analysis showed that the interactionist approach is very **scalable** as the methodology of analysis does not have to change when going from a dyadic to a multiparty interaction. The analysis is driven by the interaction's activity, and takes into account its history [14]. Previous studies show that participants in groups organize the interaction in a manner that favors one-on-one exchanges, and that conversational rules between more than 2 participants are adaptations of one-on-one ones [13, 34]. However, trust psychological models are **hardly scalable**. Indeed, when users form a group to perform a joint activity, trust is considered an emergent state of the group, and group trust assessments are more than the average of each user's trust. This means that the psychological model should change drastically since social phenomena happening during dyadic and group interactions are very different [24, 25, 29]. For instance, some of the RTS' items - such as "provides appropriate information" and "communicates well" - would need to be re-specified for situations of asymmetry during group interactions - e.g. the robot communicates properly with only one participant but not the others.

5.2 Which approach for which computational study ?

Given all the previous criteria, we provide a few guidelines on the type of computational studies for trust analysis each specific approach can tackle to answer **RQ2**. The interactionist approach is a good fit for a continuous participants' behavioral analysis throughout the interaction given its time-framing. This approach can be useful in contexts such as assistive robotics for elder care where a robot needs to adapt to different interaction modalities according to the user. It is also suited to investigate the impact of the robot's behavior on the user's response, although in a very narrow timeframe and specific interactional context, such as the user's reaction to the robot's pre-opening [33].

One thing that should be taken into account when adopting this approach is the requirements of the annotation process. First of all, the annotation requires an expert trained on the coding scheme. The teaching process can be time-consuming as the annotator has to thoroughly understand the different social concepts invoked to be able to recognize them in action. Second, even though it is conducted offline, the annotation process in itself is very time-consuming and implies a heavy mental workload. For TURIN, the annotator has to segment the video and then annotate the exhibited behaviors which requires a thorough analysis and sharp focus. Some trade-offs could be considered depending on research needs, very much like what we did in this paper on segmentation and number of annotated items for each category to alleviate the annotation process.

Given its current tools, the mentalist approach is not a good fit for real-time analysis of trust in HRI. With the important adaptations of the RTS in our study, we demonstrated the need to design a more suited analysis tool with this approach. It is best suited to study the influence of the robot's design or behavior on the user's decision to trust the robot based on an overall representation of a specific or multiple criteria relating to trust through statistical analysis. To ensure that the user's representation of mentalist models' criteria have enough time to evolve, assessments should be conducted at the beginning and end of the interaction, or at sufficiently long interval during the interaction.

6 CONCLUSION

We compared both the dominant mentalist approach coming from Psychology, and the interactionist approach coming from Interactional Sociology through a theoretical analysis and a pilot study. With this process, we showed that they differ on their orientation, generalization capability, time-framing, and scalability. Both approaches are not mutually exclusive and can complement each other: the mentalist approach can unveil differences in perceptions of the robot linked to trust variation, while the interactionist one can reveal users' trust-based behavioral changes in interaction-specific settings.

Psychology-based trust research can also involve other types of measures that are considered more "objective" through scenarios such as trust games (e.g. Prisoner's Dilemna), or other measures such as proxemics, and physiological signals. While the choice of "objective" measures should be done according to the research needs, it is important to have several different ones to avoid confounding factors when using only a single measure. Future work could involve studying the relationships between these measures and the interactionist approach.

ACKNOWLEDGMENTS

This work is supported by the Data Science and Artificial Intelligence for Digitalized Industry and Services (DSAIDIS) chair of Télécom Paris.

REFERENCES

[1] Alexander M Aroyo, Jan De Bruyne, Orian Dheu, Eduard Fosch Villaronga, Aleksei Gudkov, Holly Hoch, Steve Jones, Christoph Lutz, Henrik Sætra, and Mads Solberg. 2021. Overtrusting robots : Setting a research agenda to mitigate overtrust in automation. *Paladyn, Journal of Behavioral Robotics* 12, 1 (2021), 423–436. https://doi.org/10.1515/pjbr-2021-0029

[2] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics* 1, 1 (2009), 71–81. https://doi.org/10.1007/s12369-008-0001-3

[3] Timothy Bickmore and Justine Cassell. 2001. Relational Agents: A Model and Implementation of Building User Trust. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Seattle, Washington, USA) *(CHI '01)*. Association for Computing Machinery, New York, NY, USA, 396–403. https://doi.org/10.1145/365024.365304

[4] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[5] Olive Jean Dunn. 1964. Multiple Comparisons Using Rank Sums. *Technometrics* 6, 3 (1964), 241–252. https://doi.org/10.1080/00401706.1964.10490181

[6] Alessandro Duranti. 2005. Ethnography of speaking: toward a linguistics of the praxis. , 17–32 pages.

[7] Connor Esterwood and Lionel P. Robert. 2022. Having The Right Attitude: How Attitude Impacts Trust Repair in Human-Robot Interaction. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction (HRI 2022)*. online. https://doi.org/10.7302/3781

[8] Cecilia E. Ford, Barbara A. Fox, and Sandra A. Thompson. 1996. Practices in the construction of turns: The "TCU" revisited. *Pragmatics* 6, 3 (1996), 427–454. https://doi.org/10.1075/prag.6.3.07for

[9] Cecilia E Ford and Sandra A Thompson. 1996. Interactional Units in Conversation: Syntactic, Intonational and Pragmatic Resources. *Interaction and grammar* 13 (1996), 134.

[10] Charles O Frake. 1964. How to ask for a drink in Subanun. *American Anthropologist* 66, 6 (1964), 127–132.

[11] Goffman and Erving. 1959. Presentation of Self in Everyday Life. *American Journal of Sociology Goffman, Erving* 55 (1959), 17–25.

[12] Erving Goffman. 1981. *Forms of talk*. University of Pennsylvania Press.

[13] Charles Goodwin. 1981. Conversational organization. *Interaction between speakers and hearers* (1981).

[14] John C Heritage. 1990. Interactional accountability: a conversation analytic perspective. *Réseaux* 8, 1 (1990), 23–49.

[15] Marc Hulcelle, Giovanna Varni, Nicolas Rollet, and Chloé Clavel. 2021. TURIN: A coding system for Trust in hUman Robot INteraction. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 1–8. https://doi.org/10.1109/ACII52823.2021.9597448

[16] Hayley Hung and Daniel Gatica-Perez. 2010. Estimating cohesion in small groups using audio-visual nonverbal behavior. *IEEE Transactions on Multimedia* 12, 6 (2010), 563–575.

[17] Dinesh Babu Jayagopi, Samira Sheiki, David Klotz, Johannes Wienke, Jean Marc Odobez, Sebastien Wrede, Vasil Khalidov, Laurent Nyugen, Britta Wrede, and Daniel Gatica-Perez. 2013. The vernissage corpus: A conversational Human-Robot-Interaction dataset. *ACM/IEEE International Conference on Human-Robot Interaction* (2013), 149–150. https://doi.org/10.1109/HRI.2013.6483545

[18] Zahra Rezaei Khavas. 2021. A Review on Trust in Human-Robot Interaction. *pre-print arXiv.2105.10045* (2021). https://doi.org/10.48550/arXiv.2105.10045 arXiv:2105.10045

[19] Zahra Rezaei Khavas, S. Reza Ahmadzadeh, and Paul Robinette. 2020. Modeling Trust in Human-Robot Interaction: A Survey. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 12483 LNAI. 529–541. https://doi.org/10.1007/978-3-030-62056-1_44 arXiv:2011.04796

[20] William H. Kruskal and W. Allen Wallis. 1952. Use of Ranks in One-Criterion Variance Analysis. *J. Amer. Statist. Assoc.* 47, 260 (1952), 583–621. https://doi.org/10.1080/01621459.1952.10483441

[21] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance Human Factors. *Human Factors* 46, 1 (2004), 50–80. https:

//doi.org/10.1518/hfes.46.1.50_30392
[22] Jin Joo Lee, W. Bradley Knox, Jolie B. Wormwood, Cynthia Breazeal, and David DeSteno. 2013. Computationally modeling interpersonal trust. *Frontiers in Psychology* 4 (2013). https://doi.org/10.3389/fpsyg.2013.00893
[23] Maria Madsen and Shirley Gregor. 2000. Measuring human-computer trust. In *11th australasian conference on information systems*, Vol. 53. Citeseer, 6–8.
[24] Michelle A Marks, John E Mathieu, and Stephen J Zaccaro. 2001. A Temporally Based Framework and Taxonomy of Team Processes. *The Academy of Management Review* 26, 3 (2001), 356–376. https://doi.org/10.5465/amr.2001.4845785
[25] Richard L. Moreland. 2010. Are Dyads Really Groups? *Small Group Research* 41, 2 (2010), 251–267. https://doi.org/10.1177/1046496409358618
[26] Manisha Natarajan and Matthew Gombolay. 2020. *Effects of Anthropomorphism and Accountability on Trust in Human Robot Interaction.* Association for Computing Machinery, New York, NY, USA, 33–42.
[27] Mollik Nayyar and Alan R. Wagner. 2018. When Should a Robot Apologize? Understanding How Timing Affects Human-Robot Trust Repair. In *Social Robotics*, Shuzhi Sam Ge, John-John Cabibihan, Miguel A. Salichs, Elizabeth Broadbent, Hongsheng He, Alan R. Wagner, and Álvaro Castro-González (Eds.). Springer International Publishing, Cham, 265–274.
[28] George Psathas. 1990. *Interaction competence.* University Press of Amer.
[29] Tammy Rapp, Travis Maynard, Monique Domingo, and Elizabeth Klock. 2021. Team Emergent States: What Has Emerged in The Literature Over 20 Years. *Small Group Research* 52 (2 2021), 68–102. Issue 1.
[30] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M. Howard, and Alan R. Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. *ACM/IEEE International Conference on Human-Robot Interaction* 2016-April (2016), 101–108. https://doi.org/10.1109/HRI.2016.7451740
[31] Julian B Rotter. 1967. A new scale for the measurement of interpersonal trust. *Journal of personality* (1967).
[32] Denise M Rousseau, Sim B Sitkin, Ronald S Burt, Colin Camerer, Denise M Rousseau, and Ronald S Burt. 1998. Not So Different After All : a Cross-Discipline View of Trust. *Academy of Management Review* 23, 3 (1998), 393–404. https://doi.org/10.5465/amr.1998.926617
[33] Damien Rudaz, Karen Tatarian, Rebecca Stower, and Christian Licoppe. 2023. From Inanimate Object to Agent: Impact of Pre-Beginnings on the Emergence of Greetings with a Robot. *J. Hum.-Robot Interact.* (jan 2023). https://doi.org/10.1145/3575806 Just Accepted.
[34] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In *Studies in the Organization of Conversational Interaction*, Jim Schenkein (Ed.). Academic Press, 7–55. https://doi.org/10.1016/B978-0-12-623550-0.50008-2
[35] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. 2015. Would You Trust a (Faulty) Robot?: Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust. In *ACM/IEEE International Conference on Human-Robot Interaction*, Vol. 2015-March. 141–148. https://doi.org/10.1145/2696454.2696497
[36] Kristin E. Schaefer. 2013. The Perception and Measurement Of Human-robot Trust. *Doctoral Dissertation. University of Central Florida, Orlando* (2013). http://purl.fcla.edu/fcla/etd/CFE0004931
[37] Kristin E. Schaefer. 2016. Measuring trust in human robot interactions: Development of the "trust perception scale-HRI". *Robust Intelligence and Trust in Autonomous Systems* (1 2016), 191–218. https://doi.org/10.1007/978-1-4899-7668-0_10
[38] Samuel S Shapiro and RS Francia. 1972. An approximate analysis of variance test for normality. *Journal of the American statistical Association* 67, 337 (1972), 215–216.
[39] Tanya Stivers. 2008. Stance, Alignment, and Affiliation During Storytelling: When Nodding Is a Token of Affiliation. *Research on Language and Social Interaction* 41, 1 (2008), 31–57. https://doi.org/10.1080/08351810701691123
[40] Dag Sverre Syrdal, Kerstin Dautenhahn, Kheng Lee Koay, and Michael Leonard Walters. 2009. The Negative Attitudes towards Robots Scale and Reactions to Robot Behaviour in a Live Human-Robot Interaction Study. *Proceedings of AISB09* (2009), 109–115.
[41] Suzanne Tolmeijer, Astrid Weiss, Marc Hanheide, Felix Lindner, Thomas M. Powers, Clare Dixon, and Myrthe L. Tielman. 2020. *Taxonomy of Trust-Relevant Failures and Mitigation Strategies.* Association for Computing Machinery, New York, NY, USA, 3–12.
[42] Anqi Xu and Gregory Dudek. 2015. OPTIMo: Online Probabilistic Trust Inference Model for Asymmetric Human-Robot Collaborations. *ACM/IEEE International Conference on Human-Robot Interaction* 2015-March, 221–228. https://doi.org/10.1145/2696454.2696492