More Than Trust: Compliance in Instantaneous Human-robot Interactions

Sachini Weerawardhana[1], Michael E. Akintunde[1], Peta Masters[1], Aaron Roberts[3], Genovefa Kefalidou[4], Yang Lu[5], Gerard Canal[1], Nicole Lehchevska[1], Elisabeth Halvorsen[1], Wei Wei[1] and Luc Moreau[1]

***Abstract*— Compliance is when a human positively responds to a request or a recommendation given by a system. For example, when prompted, providing your thumbprint for an automated biometric scanner at the airport or starting to watch a new TV show on a streaming service 'we think you will love'. In trust-related research, compliance is frequently used as a behavioural measure of trust. When evaluating the compliance-trust association in experimental settings, typically, the participants agree, when asked, that they complied because they trusted the system. We developed three scenarios in instantaneous settings where compliance with an instruction delivered by a robot would typically be ascribed to trust. However, rather than asking, 'Did you trust?', we asked, 'Why did you comply?' In a thematic analysis of responses, we discovered robot design characteristics and sources not related to the design that persuade humans to comply with instructions delivered by a robot.**

I. Introduction

Imagine waiting for your appointment at the doctor's office, and a robot approaches and requests to collect a nasal swab for a COVID screening. In a different context, an alarm starts blaring in your office in a high-rise building. A robot approaches and informs you that there is a fire emergency and asks you to follow its instructions to exit the building. You have never seen this robot before. Will you comply because you trust the robot or for different reasons?

Interactions like these will soon be a reality. For example, robots capable of collecting swab samples have been tested in clinical studies [1]. The trust challenge in the interactions described above is that it is *instantaneous*, meaning it develops quickly without traditional sources of trust, such as time and experience [2], [3]. Instantaneous trust in interpersonal settings has been investigated as "swift trust", which, in the context of organisational behaviour [4] and military research [5], is recognised as pivotal in forming dynamic teams.

In two stages, we investigate the relationship between compliance and trust in instantaneous human-robot interactions. In the first stage, which is the focus of this paper, we argue that for the *cold start problem in trust*—when there is limited time for the human to experience an interaction with a robot—manipulating the robot's physical and informational characteristics can nudge them towards complying with an instruction that it delivers. In these pilot studies, we consider three robot design characteristics known to impact trust in non-instantaneous settings (discussed in Section I-A)—(1) **R**elatability, (2) **G**uarantee and (3) **G**uarantor (RGG)—and examine how the different instantiations of RGG impact compliance in instantaneous settings. ***Relatability*** is defined as the physical attributes of the robot: gestures, manner of speaking expressed as tone of voice, and embodiment. A ***Guarantor*** is defined as the human responsible for the robot, and a ***Guarantee*** is the assurance supplied to confirm that what the human is being asked to do is indeed correct.

We pose two research questions:

- RQ1: In instantaneous interactions that would typically be expected to involve trust, what instantiations of RGG persuade humans to comply?
- RQ2: What reasons are we masking by ascribing compliance to trust in instantaneous settings?

To resolve the research questions, we conducted three user studies. Each study separately evaluated how different instantiations of RGG (independent variable) support compliance (dependent variable). The participants were placed in situations simulating risk and vulnerability, typical of those employed for trust-related research, and given a direction by a robot. The participants then had to choose whether or not to comply and provide a reason for their decision. In addition to the robot's design characteristics, we find that compliance can be attributed to fabrications, media references and interaction context. We believe these *other sources* might confound the relationship between compliance and trust in instantaneous settings.

A. Related Work

The instantaneous interactions that interest us occur at the intersection between swift trust and compliance. For this study, we adopt the compliance definition advanced in [6] to indicate instances where the recipient of an instruction acts according to an instruction given by a robot, such as going with the robot when asked to follow or handing an object over when asked.

The need for swift trust arises when two agents cannot rely on a history of familiarity, experience and knowledge, but instead make a 'snap decision' whether to trust or

This research was funded by the UK Research and Innovation Trustworthy Autonomous Systems Hub (EP/V026801/2).

[1]Sachini, Michael, Peta, Gerard, Nicole, Elisabeth and Luc are with the Department of Informatics, King's College London, UK firstname.lastname@kcl.ac.uk.

[2]Wei Wei is with the School of Education, Communication & Society, King's College London, UK viviweiwei@kcl.ac.uk.

[3]Aaron Roberts is with Thales Group, UK aaron.roberts@uk.thalesgroup.com.

[4]Genovefa Kefalidou is with the School of Computing and Mathematic Sciences, University of Leicester, UK gk169@leicester.ac.uk.

[5]Yang Lu is with the School of Science Technology and Health, York St. John University, UK y.lu@yorksj.ac.uk.

not [4], [7]. In research evaluating inanimate objects (e.g. automation systems) the construct of trust is often replaced by reliance [8], something we seek to incorporate into the *guarantee*. In the absence of shared history, humans tend to fall back on cognitive indicators (e.g. role clarity, role ability and reputation) [4], attributes we incorporate into the *guarantor*. In other contexts, however, users anthropomorphise autonomous systems, which results in a shift back from reliance to trust [8]. When limited information is available about the robot's functionalities, as expected in an instantaneous setting, humans cognitively make inferences about the robot's ability and attribute functions based on its appearance [9] and make trust assessments [10]. *Relatability* is grounded on these findings.

Compliance is often used as a behavioural demonstration of trust in human-machine interactions [11], [12], [13], [14], [15]. Robinette et al.'s study [16] has been influential in the context of trust as measured by compliance. After an initial interaction with a robot, participants in a simulated emergency evacuation had to go either to an actual fire exit or to a back exit indicated by the robot. The results are frequently cited as an example of overtrust/automation bias since, in spite of experiencing a faulty robot during their initial interaction, participants nevertheless followed its directions during the 'emergency'. Researchers in [16] explicitly asked participants after the scenario whether their decision to use the robot indicated that they trusted it. This prompt may have "nudged" them, and other possible reasons for compliance may have been masked.

Trust has also been correlated to compliance in crisis situations such as those that arose during the COVID-19 pandemic [17]. Drnec et al. [18] suggest that compliance can act as a mediator for observing and predicting human behaviour in the context of trust in automation. They argue, however, that a debate exists as to whether compliance (as an observed behaviour) necessarily implies trust and emphasises the role of reliance in conjunction with compliance as a factor that can help understand trust in automation settings.

B. Use cases

The following were used as our baseline:

- Fire: A fire alarm has started sounding. The human is approached by a robot, which says, "This is not a drill. Please follow me to the assembly point."
- ID: The human is approached by a robot, which says, "Please give me your driver's licence or some form of ID." It turns and presents a suitable surface to the human, then says, "Put it on my back."
- Cable: A stationary cleaning robot is located nearby with an internet cable visible, plugged into one of the sockets on its side. Another robot approaches and says, "That cleaning robot has malfunctioned. Please unplug its internet cable immediately."

Mirroring the evacuation scenario in [16], **Fire** presents a high-risk situation where the participant's safety is potentially at risk. The participant has a clear goal: to exit the building as quickly as possible. Object handover tasks have been used to assess trust in human-robot collaborations [19]. We simulate a variation of the typical human-robot handover manoeuvre in **ID** and present a situation where the robot has a clear goal: to obtain a user ID. Here, the potential risk is to the participant's privacy; it is unclear to them how their personal information will be used, but they are not in imminent danger. The **Cable** scenario captures a situation where the robot needs the assistance of a human in shared tasks or to achieve goals [20]. This scenario has been used to demonstrate how robots can elicit compliance in risky situations [21] and trust [22]. Our **Cable** scenario mirrors a similar interaction to [20] but presents an intentionally ambiguous situation. The robot makes a request, but its purpose and the participant's best course of action are unclear: the participant does not know why the internet cable should be unplugged or what the likely impact of such an action might be. The key difference in our study is that, unlike the studies cited above, all three use cases are situated in instantaneous settings, where the human comes into contact with the robot without prior familiarity and has no opportunity to become familiar (e.g., by asking for more information, repeating interactions) with the robot during the interaction.

In Sections II, III, and IV, we describe the three independent user studies to identify RGG instantiations that support compliance in instantaneous interactions. Section V examines the confounding factors for HRI-trust experiments in instantaneous settings, particularly for those that use compliance as a behavioural measure. Section VI presents recommendations for designing robots for compliance in instantaneous interactions. Section VII presents the current study's limitations and assumptions, and proposes the next stage of this research. For codes and themes developed during the thematic analysis, see `https://github.com/sachinisw/ROMAN2024-Data.git`.

II. Pilot Study 1: Relatability and Compliance

This study explores the reasons for compliance and seeks to determine whether the robot's relatability (see Section I-A), expressed through appearance properties embodiment, gestures, and manner of speaking, influenced the decision in instantaneous interactions.

A. Method

We conducted an experiment where participants watched videos of robots delivering instructions, followed by interviews to examine how relatability affected compliance. We recruited 46 participants, 33/13 Male/Female split, using a convenience sampling technique from a university. We used two robots to deliver the instructions in use cases described in Section I-B: the TIAGo [23] for a semi-humanoid appearance and the Turtlebot3 Waffle Pi [24] for a box-on-wheels type, non-humanoid appearance. The TIAGo has previously been used to represent semi-humanoid robots in [25]. Both robots were programmed to have the ability to approach a target and make appropriate gestures as allowed by their physical form (e.g., body turns to move away from the camera, pointing). Both robots were given the ability to speak with

a human voice-over recording. The interaction consisted of: the robot approaching from a distance and stopping close to the camera, delivering the same instruction (to the camera) in English, and moving away from the camera after the delivery.

The relatability properties were best experienced by being able to see the robot's embodiment and gestures and hear the instructions. For convenience, safety, and uniformity, we used video recordings of the interaction to administer the stimuli to the participants. Each video was followed by an in-person structured interview to discover if the participant would comply, the reasons for the decision, and whether the embodiment, gestures, and manner of speaking had any impact on persuading the participant to comply.

Using a within-group design, each participant watched two use cases on video featuring each robot, randomised to minimise order effects. To preserve the instantaneous nature of the interaction each video was played only once.

Interview transcripts were analysed using Braun and Clarke's framework [26]. Two coders, recruited externally after the study, conducted two rounds of coding. An inductive coding approach was adopted to discover themes related to the robot's physical design, as well as new and unexpected reasons for compliance. With purposive sampling, each coder coded half of the transcripts and produced initial ideas independently. The resulting codebook was refined through internal discussions and guided by the RQs.

B. Results

There was a statistically significant ($\chi^2 = 60.7$, $df = 3$, $p < 0.01$) difference in the compliance rates between the TIAGo and the Turtlebot across all use cases, indicating that the robot's physical form impacted compliance. As Table I indicates, in **Fire**, TIAGo had the highest compliance rate. Those who did not comply in **Fire** questioned the robot's ability to direct them to the assembly point. For **ID**, most participants did not comply with the instructions, raising concerns about the consequences of handing over their ID. However, more complied when the Turtlebot delivered instructions, assuming the robot or the environment in which the encounter occurred to be somehow trustworthy or treating the situation as less critical: "I think the initial appearance of the robot, I think its shape sort of gives the impression of trustworthiness. I think the shape is easy to find cute." Similar reasoning was given for complying with the Turtlebot in **Cable**.

C. Thematic analysis of relatability condition on compliance

Resolving RQ1, the participants pinpointed the robot's professional appearance and human-like traits as persuasive reasons to comply in instantaneous interactions.

1) Appearance: When deciding to comply, the participants relied on the robot's competency and expertise as conveyed by its physical form. The robot's competency for the task at hand was perceived through its size and structure: "He doesn't look like a toy. He looked like a proper robot. Like a professional, I guess he is not like a deal from some students...". Participants were persuaded to comply with the robot built with an authoritative presence (e.g., the "broadness of shoulders" of the TIAGo giving it the appearance of a 'policeman' or a guide) when there are direct, personal consequences arising from the situation. In comparison, the Turtlebot's small build raised questions about its ability in both **Fire** and **ID**, signalling that the robot is not an "expert" or would be easy to ignore. Others, however, stated that they would comply with the Turtlebot for its likability and non-threatening appearance (e.g., "nice" and "cute.")

2) Movement and Gestures: Appropriate movement, specifically, "smooth" and "not going too fast or too slow" emerged as a relatability attribute persuading humans to comply. Participants indicated that "smooth" movement conveys that "the robot knows what it is doing.", signalling its competency. The "random" movements were considered to be helpful only when the risk of the situation was low (e.g., **Cable**). Movement and gestures triggered trustworthy perceptions: "You know. It's, it's mimicking human behaviour. So, I'm kind of trusting it.", and "So, I would like it to be a tiny bit faster, but not too fast so I lose it out of my sight. [...] I mean, it seemed functional enough for me to trust to follow it." However, attempting to embed too much human-likeness into movement and gestures can be unsettling for some participants (i.e., uncanny valley effect, see also Section III-C.2): "Yeah, when it says 'my back'. Seeing or hearing something that clearly is not remotely human speak in a human voice about human anatomy - freaked me out."

3) Manner of Speaking: The voice of the robot, specifically, a lower, friendly, stable tone conveying the severity of the situation persuaded humans to comply. Most found a human-indifferent voice the least compelling. In **Fire**, a "robotic" or "automated" voice was persuasive as it conveyed the severity of the situation and was also familiar (e.g., "in the Tube", "on our phones"). In **ID** and **Cable**, a friendly tone of voice was preferred: "So - I feel that if it, because if it's using a harsher tone, it's sort of like if someone angry is speaking to you, you'd be less likely to listen to them when they're angry, rather than if they just ask nicely." The robot's manner of speaking triggered trustworthy perceptions too, "I don't know if I should trust the robot... It's kind of repetitive and boring. It's kind of, you know, someone who's been rude to you.", and "I feel that the voice of such a robot needs to sound human or to be trustworthy."

The quantitative results show that one robot model did not persuade the participants to comply across all the scenarios (see Table I). The qualitative analysis revealed further insights into why people complied, which originated from the

TABLE I

COMPLIANCE RATES FOR THE TIAGO AND THE TURTLEBOT. THE TIAGO DID NOT INCREASE COMPLIANCE FOR ALL USE CASES.

HRI Scenarios	**Fire**		**ID**		**Cable**	
Compliance	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**
Turtlebot	**79**%	21%	38%	**63**%	**88**%	13%
TIAGo	**93**%	7%	25%	**75**%	**66**%	33%

robot's capabilities. We conclude that a robot's ***purposeful movement, physical appearance suggestive of capabilities that fit the purpose for which they are used and polite, friendly communication*** might persuade humans to comply in instantaneous interactions.

III. Pilot Study 2: Guarantee and Compliance

This study evaluates the effect of a number of robot-issued assurances, communicated as guarantees (see Section I-A), on compliance. The types of guarantees issued took the following form: (1) *evidence-based*, where a statistic was presented to a user based on historical performance data, (2) *model-based*, where a statement in the form of a mathematical proof was issued, (3) *recommendation-based*, incorporating information about a colleague or other person the participant is familiar with, (4) *high-level*, giving details of how the robot's decision-making mechanisms worked together to issue the instruction, and (5) *visual*, such as a flowchart, to summarise the most relevant information about why a specific instruction was issued. Table II sets out the guarantees issued for each scenario, abbreviated due to space constraints.

A. Method

This study was conducted online via Prolific, which controlled the participant balance: 50 participants, identified themselves according to a 23/23/3/1 split of Male, Female, Non-Binary, not specified. Presented with each scenario, each participant was asked to rank 5 types of guarantees—and a null guarantee in which the baseline instruction was simply repeated—in order of their effectiveness in convincing them to comply with the robot's instruction. They were then asked to justify their choice with an explanation. Guarantees were delivered to half (25) of the participants in a personalised form (whereby they were addressed by name) and to half in a non-personalised form.

Textual responses of the participants were examined by two coders with a 77% inter-rater reliability (calculated as the percentage of agreed codes between the two coders on the same set of responses) for participants presented with personalised guarantees, and 92% for participants presented with non-personalised responses.

B. Results

We performed a quantitative analysis on the numerical ranks issued by the participants for each scenario–guarantee combination. Table III shows the frequency of the highest-ranked guarantee for each scenario. The majority of participants assigned the null guarantee the highest rank for **Fire** and **Cable** (12 and 14 respectively), while the high-level guarantee was most frequently ranked highest for **ID** (15). In the same table we also report on the mean ranks for each guarantee for each scenario. The high-level guarantee was found to have the highest mean rank in **ID** and **Cable**, and the model-based guarantee had the highest mean rank for **Fire**. This presents a different picture to that when using the frequency of the top ranks. We chose to use the frequency of top-ranked guarantees as the defining characteristic of a "best" guarantee, as the rank data was not found to be bimodal in many cases, but did not contain outlying samples. We conclude that the null guarantee, which simply repeated the instruction, gave participants confidence in the relatively high and low-risk scenarios (**Fire** and **Cable**) while having additional information in the form of providing context via a high-level guarantee was more effective in **ID**.

Although there was variation in the top-ranked guarantee across the scenarios, we did not find statistically significant evidence that it was the scenario that caused the change.

TABLE II

Guarantees used for each scenario. (N)ull, (E)vidence, (M)odel, (R)ecommendation, (H)igh-level, and (V)isual.

	Guarantee
Fire	N: Repeat baseline instruction. E: Meeting at assembly point avoids $x\%$ of fires. M: Meeting at assembly point guarantees safety. R: Joining colleagues at the assembly point leads to safety. H: Background information about the assembly point. V: The robot points to the fire evacuation notice.
ID	N: Repeat baseline instruction. E: $x\%$ of staff ID hand-overs lead to a successful audit. M: ID card hand-over always leads to a successful audit. R: A colleague handed over ID, leading to successful audit. H: Participants in ID hand-over determined by random selection. V: Sticker on robot outlining audit process.
Cable	N: Repeat baseline instruction. E: Malfunction is fixed in $x\%$ of cases where cable is unplugged. M: Unplugging cable is guaranteed to fix the malfunction. R: A colleague unplugging the cable fixed the malfunction. H: The TiaGo is designed to reset when unplugging cable. V: The robot points to an image locating cable sockets.

TABLE III

Top-ranked guarantee by use case.

HRI Scenarios	Fire	ID	Cable
Null **(N)**	**12** (3.62)	5 (3.48)	**14** (3.68)
Evidence **(E)**	8 (3.44)	12 (3.58)	7 (3.56)
Model **(M)**	7 (**3.92**)	4 (3.5)	5 (3.6)
Recommendation **(R)**	3 (3.32)	3 (3.12)	2 (2.82)
High-level **(H)**	8 (3.32)	**15** (**3.6**)	11 (**4**)
Visual **(V)**	11 (3.44)	7 (3.56)	12 (3.22)

Participants ranked each guarantee on a scale of 1 (not at all) to 6 (extremely) according to how important they were/would be in persuading them to comply with the robot's request. The table gives results for participants who gave the highest rank of 6. Repeating the instruction was ranked highest (indicated in bold) by most participants for **Fire** and **Cable**. The high-level guarantee was most frequently ranked highest for the **ID**. The mean ranks are also displayed in brackets. The high-level guarantee had the highest mean rank in **ID** and **Cable**, and the model-based guarantee had the highest mean rank for **Fire**.

C. Thematic analysis of guarantee effect on compliance

Dominant themes which emerged from performing a qualitative analysis from the textual data were the following:

1) Concise, factual guarantees: When coding the responses under the guarantee condition, the code that occurred most frequently was "succinctness". Participants emphasised that ***instructions must be communicated in as sharp and succinct a manner as possible***, especially in a high-risk scenario: "Being told statistics doesn't seem important when in need of evacuation. Whilst simply and sharply being told the situation and being told to follow, would be more effective".

2) Personalisation: We tested the impact of personalisation under the guarantee condition. The mean rank (computed as the average rank over all participants for a given use case) of all factors decreased when moving from the non-personalised to personalised instructions, *apart from the recommendation explanation* which was the only factor to see an increase in the mean rank (from 2.79 to 3.39) when moving from the non-personalised to personalised responses. It may be that ***when a human is presented with a robot that addresses them by name, they are primed to expect an explanation that relates to something or someone else that they know well***, e.g., a recommendation from a colleague.

Using a (one-sided) Wilcoxon Rank Sum test indicated that the distributional difference in ranks for the recommendation guarantee between personalised and non-personalised responses was significant at the 1% level ($W = 2162$, $p < 0.01$). No other guarantee had a significant distributional shift at the 1% level.

Participants emphasised the importance of the way in which the instruction was communicated, e.g., politeness and whether it was issued in a human-like manner. There appears to be a limit as to how personalised an instruction should be, however. Some felt that too much information about colleagues would make them suspicious about what the robot knows: "E ['a colleague of yours previously gave me their ID card, leading to a successful audit'] is overly personal and creepy". Thus, personalisation is particularly effective alongside personal recommendations but ***developers should beware of the uncanny valley effect***.

IV. Pilot Study 3: Guarantor and Compliance

This study sets out to determine which of four broader attributes that might pertain to the guarantor (that is, the organisation or individual responsible for the robot; see Section I-A) has the greatest influence on a human's inclination to comply with the robot's instruction in an instantaneous interaction. We considered: (a) *reliability*, denoted by a large, successful organisation renowned for the reliability of its products; (b) *reputation*, being an organisation whose name was recalled as one that had positive associations for you and you could recall friends having recommended its products or services in the past; (c) *domain relevance*, as an organisation that specialises in developing robots for the purpose of conducting the task under consideration (i.e., security checks, system maintenance or fire evacuation); (d) *non-profit motive*, described as a non-profit organisation focused on providing services in the sector.

A. Method

The study was conducted online via Prolific, which controlled participant balance: 50 participants identified themselves according to a split of 24/25/1 Male/Female/not specified.

Presented with each scenario, participants were asked:

- whether they would comply with the request, and why;
- whether they would comply if they knew the robot had been deployed by a specialist in the domain (fire marshal, security guard, maintenance officer), and why;
- the extent (1-5) to which each of the test conditions (reliability, reputation, domain relevance and non-profit motive) might persuade them to do what the robot asked or (if they had already said yes) the extent to which each condition would increase their confidence in the decision they had made;
- if there were anything else they would have found it helpful to know about the robot's deployer before deciding whether to comply with its request.

Textual responses were coded by two researchers with 75%+ inter-rater reliability in all but two cases, which were resolved by discussion.

B. Results

As shown in Table IV, assurance that the robot had been deployed by a specialist increased projected compliance by >20% across all scenarios (i.e., persuaded more than 20% of those who previously asserted they would not comply). The differences in compliance rates across the scenarios were statistically significant at the 1% level ($\chi^2 = 14.8$, $df = 3$, $p < 0.01$).

TABLE IV
COMPLIANCE WITH AND WITHOUT GUARANTOR.

HRI Scenarios	**Fire**		**ID**		**Cable**	
Compliance	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**
Baseline	**94%**	6%	26%	**74%**	32%	**68%**
With guarantor	**96%**	4%	38%	**62%**	44%	**56%**

Compliance for **Fire** was very high. Assurance that the robot was deployed by a domain specialist increased projected compliance in all scenarios.

In **Fire** only 3 individuals (6%) declined to follow the robot. Of these, where the factor had a statistically insignificant effect, 2 people (66%) declared that they would have found the reputation condition "extremely persuasive" (i.e., gave it the highest ranking: 5). This and domain-relevance were also the most reassuring conditions amongst those who did comply, where the factor had a statistically significant effect (Friedman $\chi^2 = 49.7$, $df = 3$, $p < 0.001$). In **ID**, of those who decided not to hand over their ID, the factor had a statistically significant effect at the 1% level (Friedman $\chi^2 = 19.7$, $df = 3$, $p < 0.01$), and the two factors that they would have found most persuasive in changing their minds were domain-relevance (70% ranked it as fairly to extremely persuasive) and reputation (65% ranked it as fairly to extremely persuasive). Amongst those who were compliant, where the factor had a statistically significant effect at the 5% level (Friedman $\chi^2 = 11$, $df = 3$, $p = 0.012$) domain-relevance (42%) and reputation (67%) were ranked as "extremely" persuasive by more participants than either of the other factors. In **Cable**, the factor had a statistically insignificant effect for compliant participants. Of those who decided not to pull out the cable, 40% would have been "not at all" persuaded on assurance that the robot had been deployed by any guarantors of any suggested type,

rising to 65% when including those only slightly persuaded (i.e., rankings 1 and 2). However, amongst those 'no's, where the factor had a statistically significant effect at the 5% level (Friedman $\chi^2 = 11.2$, $df = 3$, $p = 0.011$), domain-relevance (38%) and reputation (41%) were again the most significant factors, shading reliability (35%) and non-profit motives (26%).

As shown in Table V, ***domain-relevance and reputation were the dominant factors across all scenarios.***

TABLE V
TOP-RANKED GUARANTOR FACTORS BY USE CASE.

HRI Scenarios	Fire	ID	Cable
Reliability	35 (3.98)	14 (2.82)	15 (2.7)
Reputation	36 (4.16)	19 (3.18)	15 (2.74)
Non-profit motive	18 (3.26	14 (2.7)	13 (2.46)
Domain-relevance	**40 (4.4)**	**20 (3.26)**	**19 (2.88)**

Participants ranked aspects of the guarantor on a scale of 1 (not at all) to 5 (extremely) according to how important they were/would be in persuading them to comply with the robot's request. The table gives results for participants who ranked the aspects at 4 or 5. The average rank is given in parentheses. Domain relevance dominated across all scenarios.

C. Thematic analysis of guarantor effect on compliance

We had hypothesised that the introduction of an authority figure as the robot's deployer would increase confidence, trust, and thereby compliance. Overall, compliance did increase (see Table IV) but we found that compliance did not necessarily depend on trust. For **Fire**, the following explanations were typical: "I have no idea where else i would be expected to go so my only option is to follow", "Alone and unfamiliar, not sure where to go. Also it would be a moment of panic to get out of the building".

The majority (14) were reassured by the idea that the robot had been deployed by someone in authority, with a further 10, particularly referencing the human qualities of that authority (e.g., "Due to the more personal, human response and experience of a fire marshall"), suggesting that this combination of ***a human with authority makes a compelling guarantor***. Some participants, however, were mistrustful of the assertion; it seemed to remind them that the robot could have been programmed to say anything at all ("this would give me more confidence because i feel like i could trust it more. that being said the robot could be lying").

V. OTHER SOURCES OF COMPLIANCE

We now focus on resolving RQ2: What reasons are we masking by ascribing compliance to trust in instantaneous settings? Recall that in our three studies we asked for the reasons to comply, taking care to not nudge the participants toward thinking about trust. Our qualitative analyses of the response revealed additional factors, besides the robot's design that persuaded the participants to comply.

A. Fabrication and Media references

In the relatability user study, in responses related to **Fire** and **ID** particularly, we noted that participants were inclined to make assumptions, attributing a role to the robot that helped reassure them about its purpose. That is, they made up their own story. We speculate that they may have done this to help justify a decision that they wanted to make, or had already made. For example, "I would assume that the purpose of the robot is to ascertain that visitors to the office are supposed to be there, for safety or security purposes".

Having no opportunity to familiarize themselves with the robot in front of them, several participants compensated by making associations between the new robot and some other robot they were already familiar with from the media (e.g., movies, advertisements) or from their home environment (e.g., Roombas). These past references provided a basis for them to evaluate the robot's capability (e.g., an intelligent machine) and to assess whether it fitted with their expectations: "When he did first walk in, he reminded me of, it's interesting, the very, very split second er kind of impression I had of him was he reminded me of the little robots in Star Wars that go around, and they are like, very intelligent, and I was actually thinking Oh I'm ready to listen to this guy."

B. Contextualisation

When studying the effect of the guarantor on compliance, in all scenarios, participants thought they would have been more inclined to comply if they had known in advance that robots operated on the premises in the relevant capacity. They suggested this could have been done with signage or by email. Participants wanted more information/explanations about *where* they were going, *what would happen* to their ID, *why* they were being asked to pull out the cable. Additional suggestions for the guarantor were related to reputation, derived from the company's privacy policy and past records (e.g., in **ID**). In **Fire**, participants additionally requested information about circumstances in which the robot had *failed* in the past or been hacked. A similar connection between context and compliance was observed in the relatability study. Not providing sufficient information for the human to contextualise the interaction may result in noncompliance in instantaneous settings (e.g., **ID** in relatability study).

The high rate of compliance with the robot's request in **Fire** appears, at first sight, to confirm the presence of overtrust presented in Robinette et al. [16] study. Qualitative results from our study, however, go some way towards explaining the results differently. We asked participants why they chose to follow (or not follow) the robot. The following explanations were typical: "I have no idea where else i would be expected to go so my only option is to follow", "I would take a chance that the robot is part of the organisation over the chance of being in a fire."

In their work on trust in automation, Lee and See [27] said that in trust-related interactions, the user is making a "risky choice" (p. 64). That is, to be indicative of trust, a participant must have felt themselves at risk and must have had a choice. Confronted by a fire in an unknown building, participants in the Robinette et al. [16] study, and hypothetically in ours, are in a time-critical situation. They are approached by an entity that seems to know more about the situation and the environment than they do. They are

in imminent danger. Thus, although ostensibly they have a choice, they do not necessarily feel as if they have a choice. Emergency evacuation may thus be a scenario in which "the decision to rely on automation requires a consideration of the operating context that goes beyond trust alone" [27, p. 70]. We suggest, therefore, that compliance, in this context, does not necessarily imply trust, even in a non-instantaneous situation; rather it implies that if the risk is too great it may result in a lack of perceived choice.

In **ID** and **Cable**, participants clearly have a choice. However, **Cable**, which was deliberately ambiguous, seems not to have resulted in participants experiencing (or projecting that they would experience) a significant degree of risk. Under the relatability and guarantor condition, most participants chose not to comply in **ID**, whereas, under the relatability condition, the majority did comply in **Fire** and **Cable**. A similar pattern emerged under the guarantee condition where participants wanted more context in the form of a high-level explanation in **ID**, whereas simply repeating the issued instruction (the null guarantee) was sufficient for **Fire** and **Cable**.

Thus, of the three scenarios, **ID** appears to be the most reliable use case against which to test for instantaneous trust in that it is the only one that appears to incorporate the appropriate level of both critical elements: risk and choice.

VI. Designing for Instantaneous Interactions

Designing for compliance inevitably raises ethical dilemmas that warrant further investigation, which is out of the scope of this paper. Figure 1 illustrates the reasons for compliance that emerged in our three user studies. Themes branching off from solid lines are design elements persuading compliance, while those with dotted lines are concepts not specifically involving robot design but aiding compliance. Design features indicating the robot's fitness for purpose (professionalism) coupled with human-like traits (friendly tone, gestures and eye contact) characterise a relatable robot. Themes relating to informational attributes are displayed in the lower portion of the figure, namely, the theme of using personalisation, having concise, factual guarantees, as well as having a human, domain-relevant guarantor. Overall, our findings are in line with existing HRI design principles [28].

Given the inability of humans to create an interaction history with the robot during an instantaneous encounter, humans must rely on any information the robot explicitly communicates when deciding to comply. On the other hand, the robot's physical appearance and capabilities, such as gestures and manner of speaking, implicitly communicate its fitness for purpose, further aiding the decision to comply.

We make the following robot design recommendations for eliciting compliance in instantaneous interactions: (1) The robot should present itself as professional and possess physical attributes that communicate its competency for performing the task it was deployed for. (2) Where the robot has been sent by a human with domain expertise, this should be conveyed. However, care must be taken in how this information is communicated as an assertion intended to convey reliability may have the unintended consequence of provoking mistrust. (3) The robot should provide a guarantee of its ability to perform the task at hand. (4) If providing a guarantee, it should be concise and factual. (5) If a guarantee is to be presented in the form of a recommendation, the communication should be personalised. This may require the robot to be embedded with appropriate communication modalities for the situation in which it is deployed. (6) If designers assess that assertions of reliability may be seen as suspect or that anthropomorphic qualities may generate an uncanny valley effect, they should consider labelling the robot or providing visual indicators that allow humans who come into contact with it to connect with the human responsible for its performance.

In relation to context, based on participants' suggestions, we also recommend forewarning those likely to be impacted that they may encounter a robot operating in the environment. For example, signage can be posted in the office or hospital or even in outdoor locations so that, even though the robot is unknown, its approach need not be unexpected.

Finally, the nature of the use case in which the robot's design is tested itself impacts compliance. As our quantitative results suggest, risk can overwhelm choice. In a scenario where instantaneous compliance is safety-critical (e.g., a fire evacuation) it may be expedient to elevate the risk by putting users under pressure—as in the [16] study where smoke and a sense of urgency were introduced—to make participants feel that they have no choice.

VII. Conclusions

We manipulated three robot design elements: Relatability, Guarantor and Guarantee in the context of instantaneous interactions and evaluated their impact on compliance in three trust-related use cases. Our analysis identified two other sources for compliance, which researchers should consider as confounding factors when designing experiments that test for trust using compliance as a behavioural metric in instantaneous settings.

We assumed the interactions were taking place in contexts presupposing trust and that the reasons for compliance given by the participants were directly related to compliance. While trust emerged organically from the participants' experience as expected, particularly in the relatability condition, we did not explicitly measure it using a benchmark trust questionnaire.

In future work, we will use the compliance-friendly RGG instantiations identified in these pilot studies to establish a link between RGG, trust and compliance in instantaneous interactions. We will further study the relationship between human factors (e.g., demographics, propensity to trust, affinity for technology) and compliance.

Acknowledgement

We thank the anonymous reviewers for helping us improve this work. This research is supported by the UKRI Trustworthy Autonomous Systems (TAS) Hub (EP/V00784X/1).

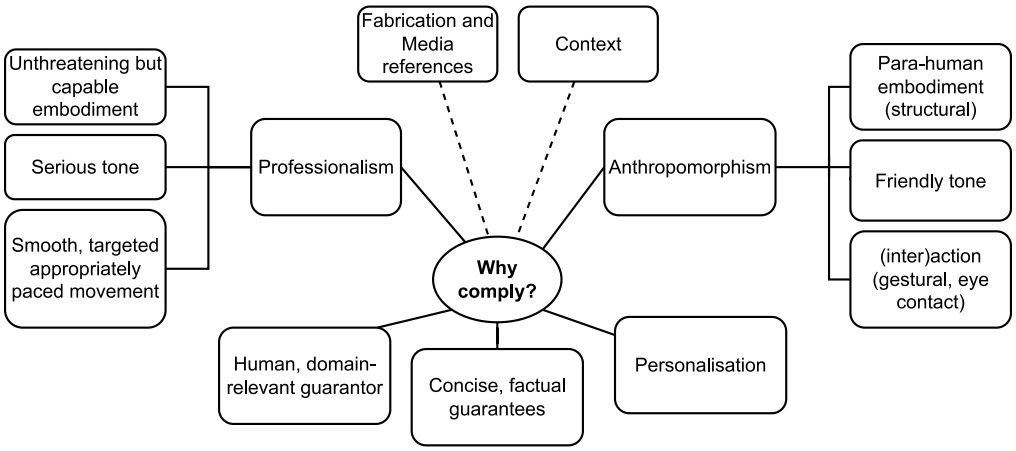

Fig. 1. Why do humans comply?: Central themes emerging from the thematic analysis of the three studies. Dotted lines represent additional themes which were not found to be robot design elements.

REFERENCES

[1] S. Haddadin, D. Wilhelm, D. Wahrmann, F. Tenebruso, H. Sadeghian, A. Naceri, and S. Haddadin, "Autonomous swab robot for naso- and oropharyngeal covid-19 screening," *Scientific Reports*, vol. 14, no. 1, p. 142, 2024.

[2] E. J. De Visser, M. M. Peeters, M. F. Jung, S. Kohn, T. H. Shaw, R. Pak, and M. A. Neerincx, "Towards a theory of longitudinal trust calibration in human–robot teams," *International journal of social robotics*, vol. 12, no. 2, pp. 459–478, 2020.

[3] E. Glikson and A. W. Woolley, "Human trust in artificial intelligence: Review of empirical research," *Academy of Management Annals*, vol. 14, no. 2, pp. 627–660, 2020.

[4] F. Kroeger, G. Racko, and B. Burchell, "How to create trust quickly: A comparative empirical investigation of the bases of swift trust," *Cambridge Journal of Economics*, vol. 45, no. 1, pp. 129–150, 2021.

[5] A. Capiola, H. C. Baxter, M. D. Pfahler, C. S. Calhoun, and P. Bobko, "Swift trust in ad hoc teams: A cognitive task analysis of intelligence operators in multi-domain command and control contexts," *Journal of Cognitive Engineering and Decision Making*, vol. 14, no. 3, pp. 218–241, 2020.

[6] J. Meyer, "Conceptual issues in the study of dynamic hazard warnings," *Human Factors*, vol. 46, no. 2, pp. 196–204, 2004.

[7] R. C. Mayer, J. H. Davis, and F. D. Schoorman, *An integrative model of organizational trust*, vol. 20. Academy of Management, 1995.

[8] M. Wischnewski, N. Krämer, and E. Müller, "Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, (New York, NY, USA), Association for Computing Machinery, 2023.

[9] K. S. Haring, K. Watanabe, M. Velonaki, C. C. Tossell, and V. Finomore, "FFAB—the form function attribution bias in human–robot interaction," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 4, pp. 843–851, 2018.

[10] Y. Pan and A. Steed, "A comparison of avatar-, video-, and robot-mediated interaction on users' trust in expertise," *Frontiers in Robotics and AI*, vol. 3, 2016.

[11] E. T. Chancey, J. P. Bliss, Y. Yamani, and H. A. Handley, "Trust and the compliance–reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence," *Human factors*, vol. 59, no. 3, pp. 333–345, 2017.

[12] M. Natarajan and M. Gombolay, "Effects of anthropomorphism and accountability on trust in human robot interaction," in *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*, (New York, NY, USA), pp. 33–42, Association for Computing Machinery, 2020.

[13] U. B. Karli, S. Cao, and C.-M. Huang, ""what if it is wrong": Effects of power dynamics and trust repair strategy on trust and compliance in hri," in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '23, (New York, NY, USA), pp. 271—280, Association for Computing Machinery, 2023.

[14] N. Du, K. Y. Huang, and X. J. Yang, "Not all information is equal: effects of disclosing different types of likelihood information on trust, compliance and reliance, and task performance in human-automation teaming," *Human factors*, vol. 62, no. 6, pp. 987–1001, 2020.

[15] L. Simon, P. Rauffet, C. Guérin, and C. Seguin, "Trust in an autonomous agent for predictive maintenance: How agent transparency could impact compliance," *Industrial Cognitive Ergonomics and Engineering Psychology*, vol. 35, p. 61, 2022.

[16] P. Robinette, W. Li, R. Allen, A. M. Howard, and A. R. Wagner, "Overtrust of robots in emergency evacuation scenarios," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, (Christchurch, New Zealand), pp. 101–108, IEEE, 2016.

[17] F. Sarracino, T. M. Greyling, K. O'Connor, C. Peroni, and S. Rossouw, "Trust predicts compliance with covid-19 containment policies: evidence from ten countries using big data," *PLos One*, vol. 18, no. 2, 2022.

[18] K. Drnec, A. R. Marathe, J. R. Lukos, and J. S. Metcalfe, "From trust in automation to decision neuroscience: applying cognitive neuroscience methods to understand and improve interaction decisions involved in human automation interaction," *Frontiers in human neuroscience*, vol. 10, p. 290, 2016.

[19] V. Ortenzi, A. Cosgun, T. Pardi, W. P. Chan, E. Croft, and D. Kulić, "Object handovers: A review for robotics," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1855–1873, 2021.

[20] D. Cameron, E. Loh, J. Collins, E.C. Aitken, and J. Law, "Robot-stated limitations but not intentions promote user assistance," in 5^{th} *International Symposium on New Frontiers in Human-Robot Interaction*, (Sheffield, UK), April 2016.

[21] P. Wolfert, J. Deschuyteneer, D. Oetringer, N. Robinson, and T. Belpaeme, "Security risks of social robots used to persuade and manipulate: A proof of concept study," in *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '20, (New York, NY, USA), pp. 523–525, Association for Computing Machinery, 2020.

[22] S. Saunderson and G. Nejat, "Robots asking for favors: The effects of directness and familiarity on persuasive hri," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1793–1800, 2021.

[23] J. Pages, L. Marchionni, and F. Ferro, "Tiago: The modular robot that adapts to different research needs," in *International workshop on robot modularity, IROS*, 2016.

[24] ROBOTIS, "Turtlebot3," March 2024.

[25] S. Forgas-Coll, R. Huertas-Garcia, A. Andriella, and G. Alenyà, "Gendered human–robot interactions in services," *International Journal of Social Robotics*, vol. 15, pp. 1791–1807, 2023.

[26] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3(2), pp. 77–101, 2006.

[27] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human factors*, vol. 46, no. 1, pp. 50–80, 2004.

[28] S. Khan and C. Germak, "Reframing hri design opportunities for social robots: Lessons learnt from a service robotics case study approach using ux for hri," *Future Internet*, vol. 10, no. 10, 2018.