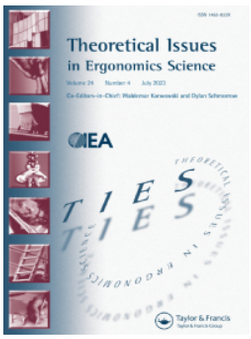

Taylor & Francis
Taylor & Francis Group

Theoretical Issues in Ergonomics Science

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ttie20

A theoretical model to explain mixed effects of trust repair strategies in autonomous systems

Richard Pak & Ericka Rovira

To cite this article: Richard Pak & Ericka Rovira (2023): A theoretical model to explain mixed effects of trust repair strategies in autonomous systems, Theoretical Issues in Ergonomics Science, DOI: 10.1080/1463922X.2023.2250424

To link to this article: https://doi.org/10.1080/1463922X.2023.2250424

Published online: 25 Aug 2023.

THEORETICAL ISSUES IN ERGONOMICS SCIENCE
https://doi.org/10.1080/1463922X.2023.2250424

Taylor & Francis
Taylor & Francis Group

REVIEW ARTICLE

A theoretical model to explain mixed effects of trust repair strategies in autonomous systems

Richard Pak[a] and Ericka Rovira[b]

[a]Department of Psychology, Clemson University, Clemson, SC, USA; [b]Department of Behavioural Sciences & Leadership, United States Military Academy, West Point, NY, USA

ABSTRACT

The topic of an autonomous system initiating trust repair has generated intense interest from researchers and has led to a stream of empirical works studying the impact of different trust repair strategies. Unfortunately, there does not seem to be a clear pattern of results or systematicity in the experimental manipulations. This is likely due to a lack of a coherent model or theoretical framework of trust repair. We present a possible theoretical model that may explain and predict how different trust repair strategies may work with different autonomous systems, in different situations, and with different people. We have adapted and applied a well-established social cognition theory that has successfully explained and predicted complex attitudinal and behavioural phenomena. The model suggests that significant variance in trust repair results may be partly due to individual differences (e.g. motivation, cognitive abilities), which have not been extensively examined in the literature, and confounded or uncontrolled study parameters (e.g. timing of trust measurement, repair frequency, workload). We hope that this theoretical model stimulates discussion toward a more theory-driven trust repair research agenda to understand basic underlying mechanisms.

ARTICLE HISTORY
Received 3 November 2022
Revised 16 August 2023

KEYWORDS
Trust repair; elaboration likelihood model; individual differences; automation; autonomy; robots; human-autonomous teams

Relevance to human factors/Relevance to ergonomics theory

There is an urgent need for a coherent model of trust repair since the current atheoretical approach is not sustainable. We present a model that may explain some existing findings in the trust repair literature, applies to most forms of autonomous systems, and makes straightforward predictions about trust repair.

1. Introduction

Successful human-autonomous system cooperation and teaming require trust (Chiou & Lee, 2023). Trust is one of the most widely studied variables in human autonomy teaming (O'Neill, McNeese, Barron, & Schelble, 2022) and has been studied in a variety of domains including but not limited to aviation, healthcare, military (Chen & Barnes, 2014), and

CONTACT Richard Pak richpak@clemson.edu

surface transportation (Zhang, Zhengming, Tian, & Duffy 2022). In a process known as trust calibration, the human's trust in the system undergoes a process of development and change as they better understand the capabilities of the system or experience new situations (Lee and See 2004). If trust is uncalibrated, that may lead to under-trust (i.e. the human distrusting a reliable system) or over-trust (i.e. the human placing too much trust in an unreliable system). A commonly used design strategy to aid trust calibration has been the use of anthropomorphism (e.g. Pak et al. 2012), or making the system appear more human. However, a complimentary strategy to manage trust may be to design the system to *behave* or *respond* in a more human way: to apologize, for example.

The notion that an autonomous system could manage and repair an individual's trust in it is not a new idea, at least in the realm of science fiction (e.g. HAL9000 from the movie *2001*). However, only fairly recently has it been given research attention (Robinette, Howard, and Wagner 2015) and a framework for its implementation (de Visser, Pak, and Shaw 2018). Since then, there has been a surge of studies that have examined the efficacy of different trust repair strategies in a wide variety of domains (for a partial review, see Esterwood and Robert 2022). We define trust repair as an act that is taken by an autonomous system, in response to a violation that decreases trust, to enhance an individual's trust (de Visser, Pak, and Shaw 2018). The violation is any act, by the autonomous system (intentional or not) that has decreased an individual's trust in it.

In a recent literature review of trust repair in human-robot interaction (HRI), Esterwood and Robert (2022) concluded that 1) there does not appear to be any kind of consistency in trust repair effects, and 2) there appears to be little theoretical guidance to try to explain and predict findings. This lack of strong theoretical guidance leads to a situation where there is an ever-increasing plethora of empirical studies examining different trust repair strategies across different autonomous system forms (e.g. AI agents, robots, autonomous cars), domains, and participant groups. If there is any consistency to be found in the current literature, it may be that the success of trust repair (at least in HRI) depends on moderators. From their review, Esterwood and Robert (2022) identified timing of the repair, violation type, and severity of violation as potential moderators.

However, there are at least two issues with the focus on the moderators of trust repair: 1) it assumes that inconsistencies in the literature can be attributed to external methodological differences that are unrelated to the trust repair strategy itself (i.e. timing of delivery, violation severity), and 2) ignores the fact that *person-related factors* (individual differences) may affect processing of the trust repair strategy. It is plausible, but currently under-examined in the current literature, that the efficacy of trust repair could be moderated by characteristics of the human in terms of experience, ability, and motivations (e.g. Gielo-Perczak and Karwowski 2003). However, this limitation is more of a criticism of the extant research than the review.

We are in total agreement with Esterwood and Robert (2022) that the lack of a strong theoretical foundation in the trust repair literature is a stumbling block because it limits our ability to explain current disparate findings, generalise results to other situations, and generate new research hypotheses. While we think that it may be premature to carry out a formal metaanalysis of general trust repair effects (partly due to the number of studies but also the wide variety of trust repair strategies studied within the available studies), it may still be useful to look for trends in the landscape of findings to try to identify inconsistencies and discern patterns in the wider trust repair literature (as Esterwood & Roberts have done within HRI). There is an urgent need for a unifying and coherent model of trust repair since, as a scientific endeavour, the current undirected, atheoretical approach is not sustainable.

de Visser, Pak, and Shaw (2018) provided a theoretical framework to describe how trust repair might work in a human-autonomy context, including mention of the possibility that individual differences may affect trust repair efficacy. Later (de Visser et al. 2020), they introduced a broader framework, incorporating the novel concept of relationship equity to explain and predict longitudinal (long term) trust. From their model, they classify actions that a robot could take to build up relationship equity into four categories: trust repair, trust dampening, transparency, and explanation. Over time, each of these actions alter the balance of relationship equity between teammates (human and autonomous system).

Building on their work but focusing on the relationship equity-building act of *trust repair*, we present a complimentary model that may explain some existing findings in the trust repair literature (which tend to look at short-term trust, not longitudinal), is applicable to most forms of autonomous systems across various domains and makes straightforward predictions about trust repair.

2. Trust repair as persuasion: applying the elaboration likelihood model of persuasion

The first key assumption in our attempt to re-interpret existing findings in trust repair is to say that the act of trust repair is essentially an attempt at persuasion; that is, the autonomous system, *via* trust repair, is attempting to change our attitude (i.e. trust) towards it. There are subtle differences in the definition of an attitude (Bohner and Dickel 2011), but for our purposes, a generally accepted definition of an *attitude* is an evaluation or judgment of a person, object, or idea that can be expressed (and measured) as an affect or how one feels about the object, cognition or what is thought about the object, and behaviour or how the attitude affects behaviour toward the object (Ajzen 2001). *Persuasion* is defined as the formation or change of an attitude in response to a message about an object (Bohner et al. 2008). Finally, Lee and See (2004) description of *trust* is particularly suitable, as they define trust as, 'the *attitude* [emphasis added] that an agent will help achieve an individual's goals in a situation characterised by uncertainty and vulnerability' (pp. 51).

When trust repair attempts are framed as attempts at persuasion, we can apply the well-accepted elaboration likelihood model (ELM; Petty and Cacioppo 1986). The ELM was itself an attempt to explain disparate findings in the early social cognition literature in persuasion and attitude change. ELM is a general theory for organising and understanding the factors and processes that influence persuasive communications (Petty & Cacioppo). The theory assumes that people are motivated to hold correct (i.e. accurate) beliefs about the world. But people do not always have the necessary resources or opportunity to vigilantly process the information content of persuasion attempts or arguments. Two factors determine whether people will be persuaded. First, does the persuasive message (i.e. trust repair attempt) contain substantive information? Second, is the person receiving the trust repair intervention motivated and have the ability (or resources) available to diligently process the message, and do they? These two factors interact to determine whether the persuasion attempt is successful at changing attitudes.

If the trust repair attempt contains substantive information about the error that caused the trust violation, and the person has the motivation and ability to process the message, and is indeed persuaded (trust is recovered), then this is the *central* route to persuasion. An example might be the trust repair strategy of *explanation* where an autonomous system attempts to explain why it failed and the person is able to understand and process

the message. The central route to persuasion is likely to be a resource-limited process (Norman and Bobrow 1975) since full processing of an information-rich trust repair strategy will be proportional to the amount of mental resources devoted to it, and thus would be highly sensitive to individual differences in abilities or motivation.

However, if the trust repair attempt does not contain any substantive information but instead makes an emotional appeal, and the person is not motivated (or does not have the cognitive resources) to independently investigate further, and trust is recovered because the person takes the apology at face value, this is the *peripheral* route to persuasion. In contrast to the central route, the peripheral route is likely to be data-limited (Norman and Bobrow 1975)—due to the paucity of information, increasing the amount of resources devoted to analysing the trust repair will have little effect (i.e. because there is little to further analyse). Thus, peripheral route trust repair is unlikely to be greatly affected by individual differences in abilities. For example, an *apology* expresses remorse but does not provide any substantive information and the human does not have the opportunity to independently investigate the error cause (e.g. distracted, lack of opportunity or resources). It is important to emphasise that whether the trust repair attempt is classified as central or peripheral is not a property of the trust repair strategy alone but is an interaction of the strategy and how the person processes it. Figure 1 summarises how we adapted ELM to explain trust repair. Note, the development and sophistication of autonomous systems are rapidly evolving so while the scope of this work focuses on systems that are more autonomous and may be able to identify failure, the theory does not rely on the autonomous system automatically realizing a failure—the system could be explicitly notified of a failure by the user.

Figure 1 also reinforces the notion that whether trust repair arises from the central or peripheral route depends on the information content of the strategy, and how it is processed by the recipient. This processing step (adjacent to the boxes labelled 'Moderators') illustrate the possible influences of individual differences in motivation, abilities (cognitive resources),

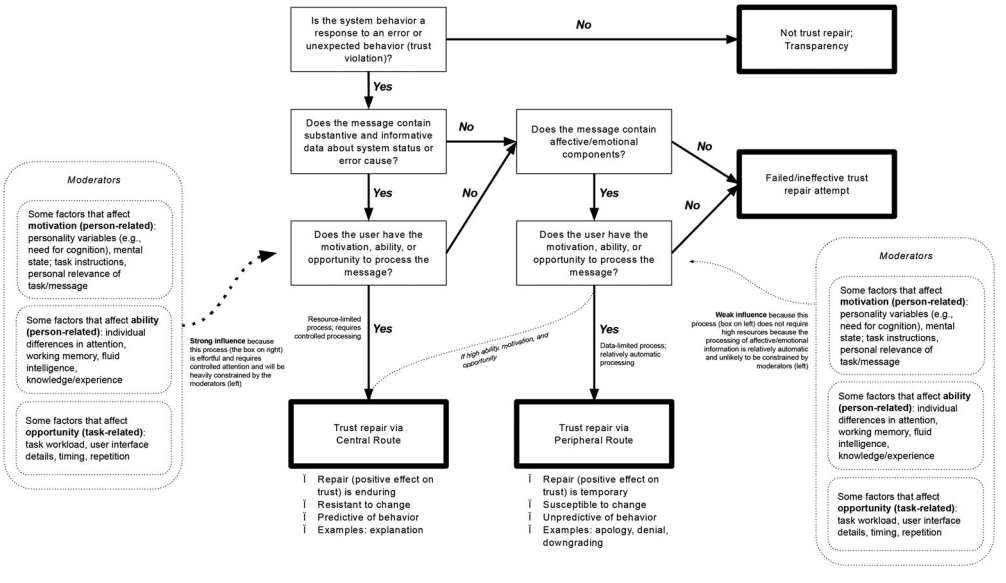

Figure 1. ELM interpretation of trust repair Interventions and effects (adapted from Petty and Cacioppo 1986).

but also task-related factors that could affect processing of the trust repair message (e.g. timing of trust repair delivery, workload, user interface details). We also theorise that violation type, which was one of the earliest identified factors influencing the efficacy of trust repair in human-human studies (e.g. (Kim et al. 2004), could be considered something that alters the mental state (e.g. attitude) of the person by altering future perceptions of subsequent actions (e.g. causal attributions).

Because of the interaction between trust repair content and recipient determines the ultimate effect, in rare circumstances, even an information-poor trust repair strategy (e.g. apology) could lead to a central route to persuasion if the person has the motivation, ability, and opportunity to self-generate potential explanations of the system's behaviour outside of the trust repair (illustrated as the dotted line above the central route in Figure 1). For example, if a self-driving car suddenly made an abrupt and unexpected movement, but apologised with no explanation, yet the driver could see that the action was taken to avoid hitting a pothole in the road, suddenly, the simplistic, affective trust repair strategy (apology) would be successful because while the apology did not offer an explanation, the driver generated their own (i.e. sensemaking, generating causal attributions). We think this may be a rare occurrence because it relies on high *motivation* of the passenger to seek out more information (e.g. need for cognition), high *ability* to have the cognitive resources (attention, working memory, reasoning ability), and an *opportunity* (low workload, clear view of the road).

This distinction between central and peripheral routes to trust repair is very useful for two reasons. First, it provides a framework to explain exactly how different kinds of moderators might influence the trust repair, or persuasion, process (e.g. situational factors, person-related individual differences). Second, the notion that trust repair can arise *via* two distinct routes is important because *how* trust repair is achieved is predicted to cause different qualities of that repair (e.g. duration, resilience). Table 1 summarises some of the major differences between central and peripheral routes to trust repair, example trust repair strategies that could take each route, as well as some general predictions about the effect of each category of trust repair that are heavily drawn from the ELM literature.

3. Reinterpreting some trust repair results through the lens of ELM

We now use this section to discuss recent results in trust repair and theorize how they could be explained within the framework of our model. As recent reviews have suggested (Esterwood and Robert 2022), comparison between studies is difficult, findings are disparate, and hence it is challenging to provide designers of AS systems with generalisable guidelines regarding trust repair strategies. Table 2 is an attempt to use our model to explain a sample of recent trust repair findings. Given the importance of the interaction between trust repair content and factors that might disrupt the processing of the repair, we coded each study along different relevant factors. Based on what was reported in each paper, we describe some of the important factors in determining which route to trust repair was used. First, we describe the study type (which may indicate general level of motivation, workload, or distraction experienced by the participants), the trust repair that was used, our estimate of the amount of information contained in that repair, our estimate of when the trust measurement was taken, the relative trust repair effect (where a '+' indicates the trust repair was successful and an '=' means trust nether increased nor decreased following the trust repair), and finally some plausible explanations for the finding using concepts from our model.

Table 1. Summary of central and peripheral routes to trust repair.

	Central Route (CR)	Peripheral Route (PR)
Definition	Trust recovery strategy that presents a high amount of relevant diagnostic or elaborative (e.g. explanatory) information content. It requires a high amount of message processing and elaboration	Trust recovery strategy that presents low relevance information content and uses simplistic cues to encourage the receiver to engage in simple inferencing. The paucity of information requires a low amount of processing
Content of repair strategy message	High in informative detail about cause of error, a logical argument that explains current or future performance. Factors that have been examined such as responsibility attribution would be considered informative detail	Low informative detail or reasoning for error; weak argument; uses affective cues or appeals to emotion, or other automatically processed social phenomena (e.g. stereotypes).
Example Trust recovery/ calibration strategies	Explanation, proficiency (with explanation), convey history of performance	Apology, denial, proficiency (without explanation, gaslighting, downgrading, any social behaviour (e.g. etiquette, anthropomorphism)
Cognitive demands / trust repair recipient's required motivation	High, effortful, controlled processing and high motivation (to understand or do well in the task)	Low, unconscious/implicit, only automatic processing is required and low motivation
General predictions about trust repair *via* this route (from empirical ELM persuasion literature)	• Trust repair will be lasting, and resistant to change • CR is more likely in low workload situations • Higher ability individuals are more likely to follow the CR • Younger adults, compared to older, are more likely to follow the CR (due to availability of processing resources) • Repetition will strengthen repair effect • High motivation will encourage CR	• Trust repair will be short-lived • PR may be relatively independent of workload • Lower ability individuals are more likely to follow PR • Older adults, compared to young, are more likely to follow the PR (due to age-related declines in available resources) • Repetition dull effect • Any factor that interrupts processing of trust violation or repair (e.g. delayed repair timing, distraction, time-pressure) is likely to encourage PR • Low motivation will encourage PR

The purpose of the table is to gain a general understanding of effects and how they may be plausibly explained by factors from our model. However, comparing the results from different studies is challenging because of wide methodological differences. First, not all studies measured trust before and after a trust violation to assess the absolute effect of trust repair. Most studies make relative comparisons of many different repair strategies by measuring and comparing their effects on trust only after a repair. Second, the time interval between violation, trust repair, and measurement varied for each study—this is noteworthy because our model suggests they affect trust repair. In addition to timing of key events, those studies that did not use between-groups designs subjected their participants to repeated trust repair attempts. Our model posits that repeated trust repair attempts will dull the effects of peripherally-induced strategies but strengthen central ones. This possibility is further discussed in the next section. Third, we do not discuss the potential moderating role of individual differences (e.g. in experience, abilities, motivations) because those variables are not reported in the studies discussed. Despite these challenges, this table is meant to be a qualitative evaluation of a small sample of existing research. It supports an understanding of the success of the trust repair strategy based on our model (amount of information content and amount of affective content) and moderators (study design characteristics and task conditions).

Table 2. Select trust repair studies, qualitative description of their parameters, and plausible explanations of findings based on our model.

Study	Study Methodology	Trust repair (TR) strategy	Estimated info. content of TR (about violation)	Timing of trust *measurement* / relative distance to TR (near/far)	Relative TR effect (+positive effect, = no effect, - negative effect; within-study)	Theorised route of TR effect (central/ peripheral), and potential model explanations or commentary
Lyons, Hamdan, and Vo (2023)	Video Scenario	Explanation	High	Before & after TR / near	+	**Central**. High information content about violation, and processing of repair. **Successful Repair**
		Acknowledgement	Low	Before & after TR / near	=	**No effect.** Low information content, no affective content, no possibility for repair
		Goal alignment	Low/med	Before & after TR / near	=	**No effect.** Low information content about violation, no affective content, no possibility for repair
Schelble et al. (2022)	Synthetic task environment; collaborating with AI	Apology	Low	End of study / far	=	**No effect**. Possible weak peripheral route effect from affective apology dissipated by end of study
		Denial	Low	End of study / far	=	**No effect**. Possible weak peripheral route effect from affective apology dissipated by end of study
Esterwood and Robert (2021)	Virtual collaborative HRI task	Explanation	High	End of study / far	+	**Central**. High information content and subject able to process repair. **Successful Repair**
		Denial	Low	End of study / far	=	**No effect**. Possible weak peripheral route effect from affective apology dissipated by end of study
		Apology	Low	End of study / far	=	**No effect**. Possible weak peripheral route effect from affective apology dissipated by end of study
Robinette, Howard, and Wagner (2015)	Virtual (online) Robot-assisted emergency evac	Information (i.e. explanation)	High	End of study / near	+	**Central.** High information content and subject able to process repair. **Successful Repair**
		Immediate apology	Low	End of study / far	=	**No effect**. Possible peripheral route effect was successful but too weak or transitory to measure at end of study
		Delayed apology	Low	End of study / near	+	**Peripheral**. Trust repair was near to time of measurement. **Successful Repair**
		Immediate promise	Low-med	End of study / far	=	**No effect**. Possible peripheral route effect was successful but too weak or transitory to measure at end of study
		Delayed promise	Low-med	End of study / near	+	**Peripheral**. Trust repair was close to time of measurement. **Successful Repair**

(*Continued*)

Table 2. Continued.

Study	Study Methodology	Trust repair (TR) strategy	Estimated info. content of TR (about violation)	Timing of trust *measurement* / relative distance to TR (near/far)	Relative TR effect (+positive effect, = no effect, - negative effect; within-study)	Theorised route of TR effect (central/ peripheral), and potential model explanations or commentary
Kohn et al. (2018)	Viewed scenarios of a self-driving vehicle	Timed apology (delayed from violation)	Low	After scenario / near	+	**Peripheral**. Message contained affective components inducing a peripheral effect relatively near to measurement time (at end of scenario near where violation occurred). **Successful Repair**
		Apology with process attribution	Low-med	After scenario / near	=	**No effect.** Weak peripheral effect may have been disrupted by presentation of extra information; or extra information may have counteracted affective effect of apology; or weak peripheral effect may have been lost by measurement time (end of scenario). Within-subject nature meant repetition of trust repairs further weakening a possible already weak peripheral effect. Trends in their data show slight but non-significant positive trust repair
		Apology with entity attribution	Low-med	After scenario / near	=	**No effect.** Weak peripheral effect may have been disrupted by presentation of extra information; or extra information may have counteracted affective effect of apology; or weak peripheral effect may have been lost by measurement time (end of scenario). Within-subject nature meant repetition of trust repairs further weakening a possible peripheral effect. Trends in their data show slight but non-significant positive trust repair
		Denial	Low	After scenario / near	=	**No effect**. This strategy (denying an error) may have been too unbelievable (the violation was undeniable) and would not have worked

(*Continued*)

Table 2. Continued.

Study	Study Methodology	Trust repair (TR) strategy	Estimated info. content of TR (about violation)	Timing of trust *measurement* / relative distance to TR (near/far)	Relative TR effect (+positive effect, = no effect, - negative effect; within-study)	Theorised route of TR effect (central/ peripheral), and potential model explanations or commentary
		Denial with process attribution	Low-med	After scenario / near	=	**No effect**. This strategy (denying an error) may have been too unbelievable (the violation was undeniable) and would not have worked
		Denial with entity attribution	Low-med	After scenario / near	=	**No effect**. This strategy (denying an error) may have been too unbelievable (the violation was undeniable) and would not have worked
		Explanation	High	After scenario / near	=	**No effect**. Explanation may not have had substantive detail of nature of violation
		Gaslighting	Low	After scenario / near	=	**No effect**. This strategy (denying an error) may have been too transparent (the violation was undeniable) and would not have worked; or weak peripheral effect may have been lost by measurement time (end of scenario). Within-subject nature meant repetition of trust repairs further weakening a possible peripheral effect. Trends in their data show slight but non-significant positive trust repair
		Assurance of control	Low-med	After scenario / near	=	**No effect**. As a strategy that contains some information (but less than explanation), and a mild affective tone (confidence), this strategy may have induced a moderate trust repair effect (midway between central and peripheral) that was too weak to measure at end of study. Within-subject nature meant repetition of trust repairs further weakening a possible peripheral effect. Trends in their data show slight but non-significant positive trust repair

In Lyons, Hamdan, and Vo (2023) researchers investigated 4 types of explanation strategies following the violation of expectations by a robot using a video-based scenario study. The different types of explanation strategies varied in the amount of information content. Indeed, if looking at this work through our model (Figure 1), one can see that the high information content found in one of the explanation strategies coupled with the scenario study methodology (likely low workload and thus allowed full attention) resulted in participants successfully processing the trust repair *via* the central route. However, the other three explanation conditions (no explanation, acknowledgement, and acknowledgement plus an explanation that the unexpected action matched the mission goals) were unsuccessful in repairing trust. As our model suggests, the other three explanation conditions did not contain substantive informative content about the error cause, nor did the repairs contain affective emotional messaging resulting in a failure to repair trust (see Table 2 for description). This research demonstrates how the central route may be used successfully to support trust repair.

In contrast, Kohn et al. (2018) found success in using the peripheral route to repair trust following faults in self-driving car video vignettes. Of the ten repair strategies investigated, researchers found that using an apology had a positive impact on trust as compared to no repair. It could be that participants accepted the apology at face value resulting in the repair being successfully processed *via* the peripheral route. However, it seems as if the apology with entity or process attributions were not as effective (non-significant) as was the simple apology. We speculate that this resulted in a worst-of-all-worlds situation with weakened peripheral route processing. We deduce that the peripheral route was weakened by the additional information provided to participants, and that the messaging did not contain sufficient informative content about the cause for the repair to encourage central route processing. The denials and gas lighting repairs had no affective content and contained insufficient informative content about the error. That, coupled with denying an undeniable violation was simply ineffective, resulting in no repair.

Robinette, Howard, and Wagner (2015) found that timing of the trust repair was essential in success but could not definitively explain why. Specifically, Robinette, Howard, and Wagner (2015) provided participants with five repair types (four with high affective content: immediate and delayed apology and an immediate and delayed promise, and a delayed explanation with high informative content). However, only the delayed repairs were successful. Our model suggests that it is due to the repair and measurement being in close proximity., The immediate trust repairs were relatively far from the trust measurement event, and hence were possibly less successful because while they may have induced a peripheral route trust repair, it dissipated by the measurement time (end of study). Peripheral route effects are not expected to be long lasting compared to central route effects. Thus, our model suggests that the relative time interval between trust repair and trust measurement may have played a large role in their results. The role of timing intervals is discussed in more detail in the next section.

Last, a recent study by Xu and Howard (2022; not shown in Table 2) found trends that showed an emotional apology, with high affective content, seemed to affect trust more than the other conditions (no repair, a baseline apology with low affective content, and an apology with low affective content plus an explanation with medium informational content). While their results were not significant, they did find an overall trend of affective strategies (apologies) affecting trust more than an informational strategy (explanation). With some caution

given to its non-significance, how would our model explain their findings? We believe this effect is due to the level of workload in the task. Xu and Howard's study, in contrast to other online studies that were passive viewing of scenarios, involved active task involvement with some parts of the study requirement the participant to drive manually, and then receiving automated driving assistance that was unreliable. This is a relatively dynamic, and high workload environment compared to other studies that merely showed synthetic videos of robots/agents and humans doing tasks. Workload may differentially affect the trust repair route because of its effects on the ability to notice or process the trust repair message. Affective trust strategies (e.g. apology) may be relatively impervious to workload because they are processed relatively automatically. However, information-rich trust repairs such as explanations are sensitive to workload because they require controlled processing resources (i.e. attention and memory). A more detailed discussion of the effect of workload on trust repair appears in the next section.

4. General predictions from our model of trust repair

The argument we have made is that the ultimate quality and effect of the trust repair is dependent on an interaction between characteristics of the trust repair strategy (information content, cognitive demand) and how it is processed by the recipient (ability, motivation, opportunity). In the previous section, we discussed select studies that supported some aspects of this assertion. However, because of each studies idiosyncrasies (e.g. confounds of some factors, no measurement or control of others), our conclusions were highly speculative, and the conclusions were meant to illustrate how our model could account for the findings. Thus, we use this section to predict how different model-related factors might affect trust repair and, in the process, propose possible future studies.

Instead of predictions about the efficacy of specific trust repair strategies, our predictions revolve around the effect of some of the moderators on trust repair that arose *via* the central or peripheral route. As a reminder, central versus peripheral routes refer to successful attitude change (persuasion), or in our case, trust repair and can be arrived at through a variety of means (Figure 1). Central and peripheral routes represent endpoints on a continuum of pathways to persuasion and we focus on the endpoints only for clarity of explanation. At the simplest level, central route trust change comes about from a trust strategy that contains a high amount of information and is deeply processed by an able recipient and peripheral route trust change comes about from low-information strategies that may contain affective information that do not require deep processing from the recipient.

4.1. Predictions about individual differences

Because of the wide potential range of individual differences that can affect processing of trust strategies, we have limited our focus to cognitive and motivational/attitudinal factors. In addition, because of the inherent interaction between the trust repair, individual differences, study methodology (survey vs. experiment), and timing (of repair and measurement), it can be confusing to make predictions. So, our predictions are general predictions of the direction of effects in an ideal situation. We begin our predictions with simple main effects, and where appropriate, discuss interesting possible interactions.

4.1.1. Effect of cognitive ability

Basic cognitive abilities such as fluid intelligence (reasoning ability), attention, and working memory are important abilities that underlie complex thought and behaviour (Burgoyne and Engle 2020). Individual differences in these abilities, especially controlled attention, can explain variance in performance in a wide variety of tasks and situations (Draheim et al. 2022). Thus, it is not unreasonable to assume that differences in these abilities would account for differences in processing and interpretation of trust repair strategies (e.g. Nayyar and Wagner 2018) that differ in information content. For the purposes of ability predictions, we will discuss general fluid cognitive ability in the latent factor level, rather than predictions about specific ability indicators (e.g. measures of attention, working memory, reasoning).

General cognitive ability is expected to play a significant role in the ability to process a trust repair. Low ability participants are not expected to have any spare resources or capacity to deeply process the information content in a trust repair strategy (if any), and thus are likely to experience peripheral route trust repair compared to central route. However, those with high ability are expected to have enough capacity and resources to process complex trust repair messages and would experience trust repair *via* both routes (Figure 2(a)). This, however, is assuming trust measurement is made immediately after the repair.

Because of our proposition that central route trust repair is more durable and long-lasting than peripherally-derived trust repair, the timing of trust measurement is important. If trust measurement is made early in the study (Figure 2(a)), it is likely to show a higher peripheral trust repair effect than if measurement is taken some time after a trust repair attempt (Figure 2(b)). After time, it is plausible that the effect of peripherally-derived trust repair has dissipated compared to the central route effect. The effect of timing is discussed in more detail further below.

4.1.2. Effect of dispositional variables: attitudes & motivation

Motivation is a psychological construct used to explain factors that cause organisms to initiate or terminate behaviour. Motivation can be classified as intrinsic or extrinsic (Ryan and Deci 2000). Intrinsic motivation is carrying out an activity for its own sake, or to satisfy

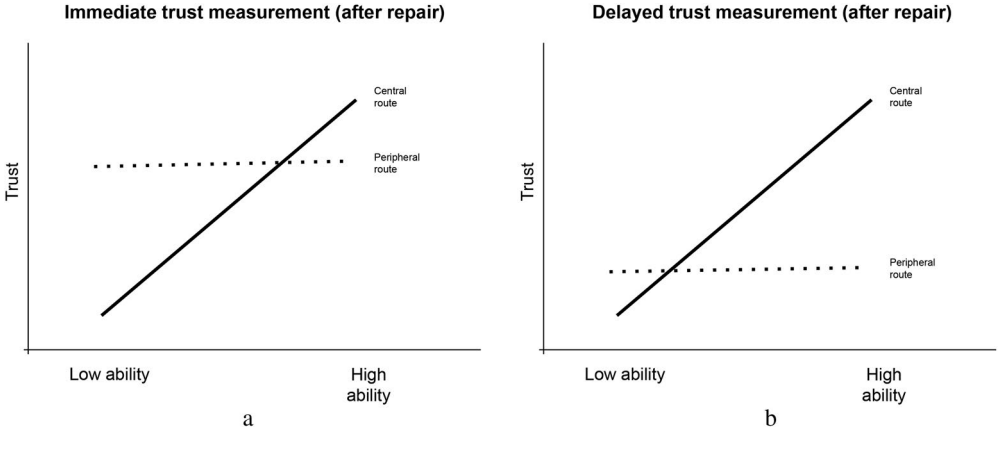

Figure 2. (a) Predicted effect of ability level on persuasion route when trust is measured immediately after repair delivery. (b) Same effect if trust is measured after a delay.

an internal need such as curiosity while extrinsic motivation is carrying out an activity for some external outcome or value such as money (Ryan & Deci). As a psychological construct, it is often measured through various means depending on the research question. In the persuasion literature, the intrinsic motivational concept of need for cognition (NFC), or an individual difference that reflects a person's desire or inclination for cognitively effortful actions (Cacioppo and Petty 1982). The correspondence between measures of NFC and behaviour were shown in studies that show that individuals who were higher in NFC were more less likely to exhibit social loafing (the phenomenon of individuals exhibiting reduced effort in a task if carried out in a group) compared to individuals who were lower in NFC. This confirmed the notion that high NFC is indicative of a willingness to engage in cognitively effortful activities even when not necessary to do so.

The role that this variable plays in persuasion has been extensively studied. Intuitively, the results show that individuals who were higher in NFC were more likely to deeply process persuasive messages and thus experience persuasion *via* the central route while low-NFC individuals were less likely to process persuasive messages (Cacioppo et al. 1986). In addition, consistent with ELM, the central-route attitude change by high-NFC individuals showed a greater correspondence with actual behaviour versus peripherally-routed attitude change.

The results with NFC show that individual differences can indeed affect the level of persuasive communications processing. Within the context of human-autonomous system interaction, relevant individual differences that may affect motivation may be dispositional attitudes about technology or autonomy such as trust propensity (Merritt et al. 2019), complacency potential (Singh, Molloy, and Parasuraman 1993), or attitudes toward specific technology such as the negative attitudes toward robots scale (NARS; (Nomura et al. 2008)). Unlike NFC, which can be easily interpreted as an unambiguous influence on motivation (i.e. NFC may directly energise an organism to seek out cognitive activities), the aforementioned variables may be less directly related to a motivation for action. However, it is possible that attitudes toward technology (e.g. NARS) may indirectly influence some aspects of motivation such as the desire to initiate or continue engaging with a robot.

4.2. Predictions related to task/Study-Factors

Because of the different qualities of trust depending on whether they are central or peripherally-derived, we expect that factors related to study and task design will play a large role in whether trust repair effects are observed.

4.2.1. Effect of workload and repair repetition

Workload (induced *via* the study methodology used) is expected to have a similar main effect as ability levels such that high workload (at trust repair attempt) is expected to result in a peripherally-induced repair effect primarily through increased thought distraction. When workload is high, a trust repair message with either low or high content will only be lightly processed (i.e. trust repair *via* the peripheral route). However, in low workload situations, centrally routed trust repair is more likely because additional cognitive resources are available to process the trust repair message. In sum, central route trust repair is likely if low workload is coupled with a high-information content message.

If low workload is paired with a low-content message, a peripheral route is likely (Figure 3(a)).

Thus, at a general level, we expect that scenario surveys are more likely to show centrally-induced trust repair compared to the same study design *via* an experiment because of the additional cognitive resources available during a scenario study where the subject is simply reading (or watching) vignettes and asked to rate the actions of the characters. This is easily testable by manipulating workload or comparing the results of two identical studies (in violation type, repair) but differing in study methodology.

The rationale that repetition may affect trust repair is that repetition simply provides more time and opportunity to scrutinise message content–in effect it may act in opposition to workload/distraction. Depending on the amount of information content, and whether initial trust repair was derived centrally or peripherally, additional scrutiny is predicted to cause greater differentiation by trust strategy information type: low-information-content trust strategies are expected to weaken trust as the recipient realises the flaws (i.e. affective repair strategy) while high-information-content trust strategies are expected to enhance trust (Figure 3(b)) as the additional time and opportunity to scrutinise the message allows the recipient to fully comprehend and understand.

4.2.2. Timing of trust measurement and timing of trust repair

According to the ELM interpretation of trust repair, there are likely to be effects on the resilience and duration of the trust repair depending on whether it was achieved *via* central or peripheral routes. Trust repair from the central route is thought to be more durable (long-lasting) and resistant to change because it came about from a more deliberative process (i.e. it provided more substantive information and required expending more cognitive resources such as attention). The greater amounts of attention and elaboration required will enhance the memory trace and make it easier to retrieve later (e.g. Craik and Lockhart 1972; Slamecka & Graf, 1978; Crocker, Fiske, and Taylor 1984). However, trust change from the peripheral route is more likely to be short-lived, and subject to change because it arose from a simpler process that does not require cognitive work (or from a process that inhibits

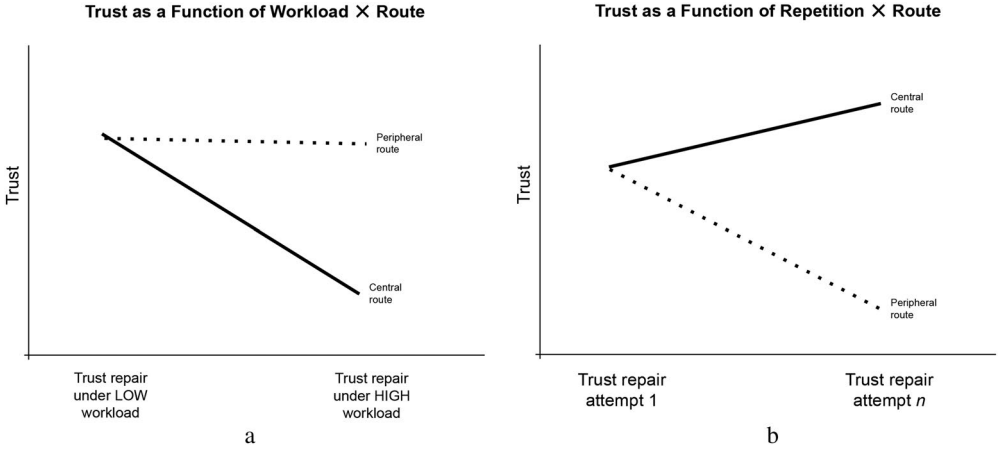

Figure 3. (a) Predicted effect of workload on trust by persuasion route. (b) Predicted effect of trust repair repetition on trust by persuasion route.

elaborative processing), such as a highly affective or emotional cue, or a simple, easy-to-understand inference. Because this memory trace is much weaker, it is less likely to be retrieved later. This is illustrated in Figure 4 which shows a decline in peripherally-derived trust repair compared to centrally-derived trust repair as a function of time.

This leads to a second prediction that depending on when the repair-induced trust measurement is made, one is likely to see different effects (Figure 4). We think this for two reasons; first, a delayed trust measurement requires the subject to recollect back to the specific violation and trust repair while the immediate measurement does not–thus placing demands on short term and working memory. Second, as discussed, we expect that peripherally-derived trust repair to be shorter-lasting than centrally-derived trust repair. If the trust measurement is made soon after the repair (trust measurement 2), there is not likely to be an observable difference between central and peripheral trust strategies. However, if the trust measurement is made later (e.g. at study conclusion; trust measurement 3), there is likely to be a difference such that central route strategies will show an enduring positive effect on trust whereas the trust repair from peripheral route strategies may have dissipated.

In addition to the timing of trust measurement, timing of the trust repair attempt has been empirically shown to affect trust repair (Nayyar and Wagner 2018; Robinette, Howard, and Wagner 2015). In several studies, delayed trust repair appears to be more effective than immediate trust repair. Our model suggests that one possible reason may be related to the short-lived properties of peripheral trust repair and a confound of timing of the repair, and the timing of the trust measurement. Immediate trust repair coincided with a trust measurement made at a later time after the repair event (end of the study) while the delayed repair was close to the trust measurement event (end of study). Presenting an immediate low-information trust repair strategy (apology) induced a peripheral trust repair effect that dissipated by the end of the study (Figure 5(a)). However, in the Robinette, Howard, and Wagner (2015) study, when the trust repair was delayed, the time interval between repair and trust measurement was reduced (Figure 5(b)). When measured that soon after the trust repair event, it is possible that the peripherally-induced trust repair effect had not yet dissipated.

In addition to being predicted to be longer in duration, the quality of trust repair is predicted to be more resistant to future trust violations when trust repair is derived *via* the central route. Figure 6 shows how trust repaired *via* a central or peripheral route is expected

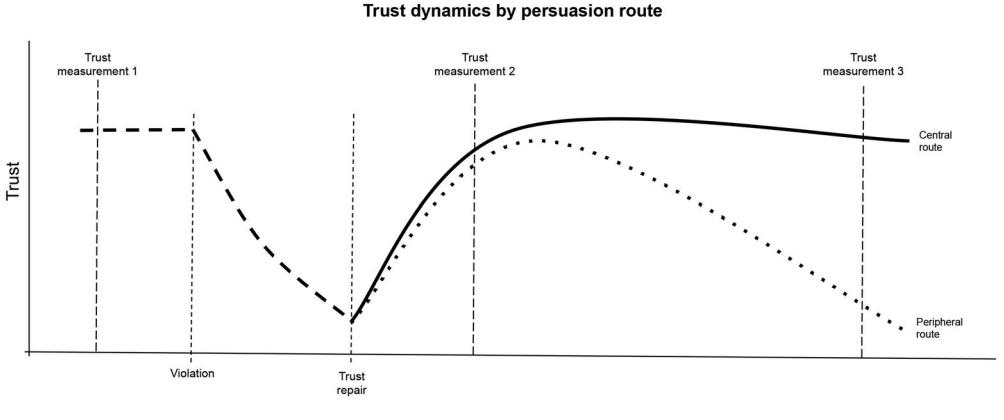

Figure 4. Trust dynamics for central and peripheral route trust. Peripheral trust repair is theorized to be less durable than central trust repair.

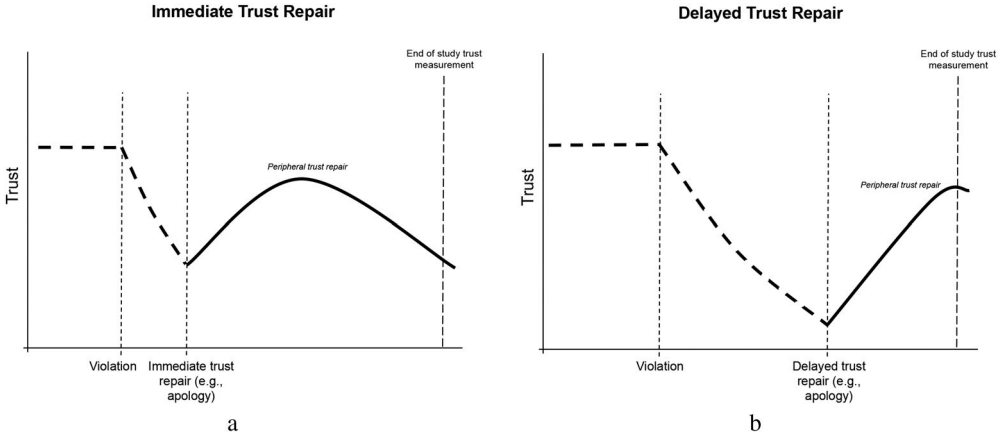

Figure 5. (a) Immediate trust repair and measurement at end of study. (b) Delayed trust repair and measurement at end of study.

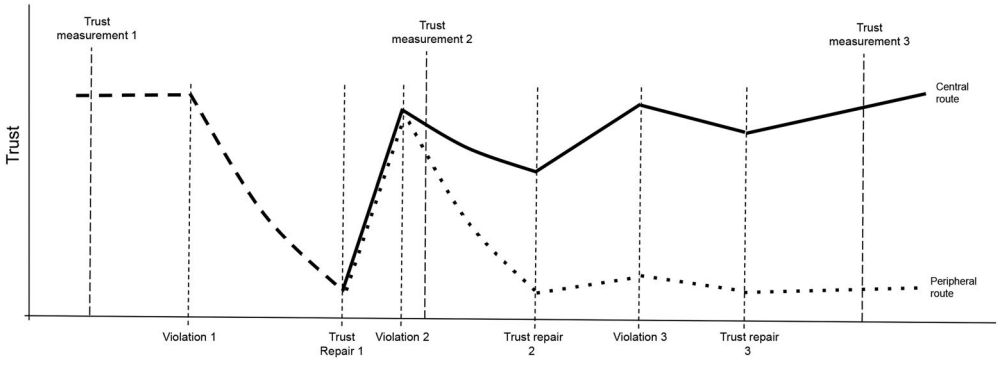

Figure 6. Trust dynamics after several trust violation and repair attempts for central and peripheral routes.

to behave with subsequent violations and trust repair attempts. Trust repair derived *via* the peripheral route is expected to be more brittle and sensitive to future violations. In the figure, this is illustrated by the dotted line. Trust repair 1 will quickly result in regained trust comparable to centrally derived trust. However, the recovery of trust will rapidly decline in comparison. After the second violation, peripherally derived trust is expected to decline more than centrally derived trust, which only declines a minimal amount due to the predicted brittleness and temporal qualities of peripherally-derived trust. Note that if trust measurement is made at this point (trust measurement 2) the differential trajectories are not likely to be seen. Subsequent violation/trust repair cycles are expected to have similar effects with greater differentiation observable between central and peripherally derived trust repair (e.g. trust measurement 3).

5. Conclusion and Recommendations for Researchers

Trust repair is one tool in the repertoire of possible system-generated behaviours that an autonomous system could use to maintain ideal trust calibration. Trust repair aims to raise

trust that has been momentarily lost due to a machine failure, not to raise trust to the highest level possible; to encourage accurate trust calibration. We stress that this interim model is based on limited data in a rapidly evolving area of research. However, the preliminary model seems to provide simple and plausible explanations for existing results and a clear theoretical foundation for the explanation and development of new trust repair strategies. The model also suggests how to quantify different trust strategies (amount of information, cognitive demands) that are more descriptive and precise than current labels (apology, explanation). The model may also explain why there may be such inconsistent findings thus far: the efficacy of trust repair is sensitive to a variety of factors that may not yet have been systematically controlled or manipulated (individual differences, repair, and trust measurement timing, study procedure). Importantly, this model is independent of the form of the autonomous system and can explain results in HRI, AI agents, as well as autonomous vehicles (Table 2).

Although this model is most applicable to explaining and predicting the effects of trust repair (after trust has declined in response to a trust violation), the two major concepts of 1) analysing the information content of the machine's message and 2) awareness of individual differences in abilities and motivations affecting the user's processing of the message may also apply to the related concepts of *transparency* and *explainability. Transparency* is the idea that, to enhance human understandability of an opaque system, the 'responsibilities, capabilities, goals, activities, and/or effects of automation should be directly observable' (Skraaning and Jamieson 2021) by the human user/teammate. Unlike trust repair, which is delivered after a trust violation to affect trust, transparency is meant to be a continuous method for keeping the user aware of the automation's state and operations to increase the user's understanding and knowledge. Transparency is typically presented with a display or message (e.g. written text, or visual indicator) which can vary in information content and complexity and may require varying levels of human ability to understand (e.g. attention, working memory). In Figure 1, our model terminates if the system's message is not a result of a system error (top right) because in that situation, trust is not expected to be harmed. However, we do not see any inherent constraints that prevent the application of this model during periods of stable trust. Thus, the extent to which transparency is achieved (i.e. comprehension) should depend on similar factors as trust repair (i.e. persuasion).

A similar concept, in artificial intelligence (AI) is that of *explainability*. Explainability is 'a human understandable explanation that expresses the rationale of the machine' (Doran, Schulz, and Besold 2017). There may be different ways to achieve explainable AI but, like transparency, the goal is to keep the human aware of the underlying rationale or logic of the actions of AI. Also, like transparency, altering trust specifically is not necessarily the goal of explainability. But the factors that affect trust repair are likely to also affect whether explainability has been achieved; to achieve explainability, the machine must present an appropriate rationale/explanation (varying in information content) which must be human understandable (depends on individual differences in abilities). Ultimately, explainability may enhance overall trust in AI. We hope that this model can be extended to explain and predict results in those areas of human-autonomy interaction.

Even as a tentative model, it shows its utility by pointing out the methodological differences between studies that may limit generalisability and emphasising factors that may

affect whether the trust repair is achieved (and measurable). In addition, the model makes specific, testable predictions. Finally, the model provides some straightforward recommendations for the design of future trust repair studies:

5.1. Trust repair manipulation

- Carefully analyse or manipulate, and report the information content or other relevant characteristics of the trust repair strategy; does it contain specific information about the cause of the trust violation?
- Assess the information processing demands of the trust repair strategy message as it may affect different routes to persuasion
- Characterise the affective content of the strategy, even if not deliberately manipulated
- Carefully consider (or manipulate) the placement of the repair in relation to the violation and measurement event.

5.2. Study and task design

The empirical procedure used to examine trust repair may be confounded with workload or cognitive demands. Surveys or scenario studies are likely to have low workload during violation or repair (as participants will simply be passively reading or viewing) while medium to high fidelity studies (with possible multi-tasking demands) with actual systems are likely to be higher in workload during the violation or trust repair.

- If the study will use a task simulation (rather than a survey), conduct a task analysis to precisely characterise the cognitive requirements and the placement of violations, trust repair, and trust measurement
- If a study uses multiple methods (e.g., scenarios and experimentation) workload should be measured to help interpret findings (e.g., NASA task load index workload measure)
- To understand the absolute, not relative, effect of trust repair, trust measurements should be made pre- and post-repair
  - Trust measurements should also be sensitive to the time course of trust repair effects and make multiple trust measurements post-repair (e.g., immediately, and at study conclusion)
  - The timing of the repair has been shown to be important for trust repair effects. Careful thought should be given to the timing of the trust repair delivery
  - Central and peripheral routes to trust repair are theorised to react differently to repetition. So, studies that use within-subjects designs (wherein subjects may receive many different trust repair attempts even if they occur in different conditions) should consider this effect when interpreting results.
  - Because of the complex attribution processes associated with trust repair, it would be useful to ask participants, at trust measurement, why they made a particular rating to understand exactly what they are rating

5.3. Participant characteristics

- Record (and report) relevant subject characteristics (cognitive abilities, dispositional/attitudinal characteristics, experience)
- Subject motivation affects message processing, so researchers should take careful consideration of sources of extrinsic motivations (e.g., task instructions to subjects; incentives), and intrinsic motivations (e.g., need for cognition)

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Richard Pak http://orcid.org/0000-0001-9145-6991
Ericka Rovira http://orcid.org/0000-0002-4820-5828

References

Ajzen, I. 2001. "Nature and Operation of Attitudes." *Annual Review of Psychology* 52 (1): 27–58. https://doi.org/10.1146/annurev.psych.52.1.27

Bohner, G., Erb Hans-Peter, Siebler Frank 2008. "Information Processing Approaches to Persuasion: Integrating Assumptions from the Dual-and Single-Processing Perspectives." In *Attitudes and Attitude Change* edited by William D. Crano, Radmila Prislin, 161–188. Psychology Press.

Bohner, G., and N. Dickel. 2011. "Attitudes and Attitude Change." *Annual Review of Psychology* 62: 391–417. https://doi.org/10.1146/annurev.psych.121208.131609

Burgoyne, A. P., and R. W. Engle. 2020. "Attention Control: A Cornerstone of Higher-Order Cognition." *Current Directions in Psychological Science* 29 (6): 624–630. https://doi.org/10.1177/0963721420969371

Cacioppo, J. T., and R. E. Petty. 1982. "The Need for Cognition." *Journal of Personality and Social Psychology* 42 (1): 116–131. https://doi.org/10.1037/0022-3514.42.1.116

Cacioppo, J. T., R. E. Petty, C. F. Kao, and R. Rodriguez. 1986. "Central and Peripheral Routes to Persuasion: An Individual Difference Perspective." *Journal of Personality and Social Psychology* 51 (5): 1032–1043. https://doi.org/10.1037/0022-3514.51.5.1032

Chen, J. Y., and M. J. Barnes. 2014. "Human–Agent Teaming for Multirobot Control: A Review of Human Factors Issues." *IEEE Transactions on Human-Machine Systems* 44 (1): 13–29. https://doi.org/10.1109/THMS.2013.2293535

Chiou, E. K., and J. D. Lee. 2023. "Trusting Automation: Designing for Responsivity and Resilience." *Human Factors* 65 (1): 137–165. https://doi.org/10.1177/00187208211009995

Craik, F. I., and R. S. Lockhart. 1972. "Levels of Processing: A Framework for Memory Research." *Journal of Verbal Learning and Verbal Behavior* 11 (6): 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

Crocker, J., S. T. Fiske, and S. E. Taylor. 1984. "Schematic Bases of Belief Change." In *Attitudinal Judgment (J.R. Eiser, Ed.)*, 197–226. New York, NY: Springer.

de Visser, E. J., R. Pak, and T. H. Shaw. 2018. "From "Automation" to "Autonomy": The Importance of Trust Repair in Human-Machine Interaction." *Ergonomics* 61 (10): 1409–1427. https://doi.org/10.1080/00140139.2018.1457725

de Visser, E. J., M. M. M. Peeters, M. F. Jung, S. Kohn, T. H. Shaw, R. Pak, and M. A. Neerincx. 2020. "Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams." *International Journal of Social Robotics* 12 (2): 459–478. https://doi.org/10.1007/s12369-019-00596-x

Doran, D., S. Schulz, and T. R. Besold. 2017. *What Does Explainable AI Really Mean? A New Conceptualization of Perspectives*. arXiv [cs.AI]. arXiv. http://arxiv.org/abs/1710.00794

Draheim, C., R. Pak, A. A. Draheim, and R. W. Engle. 2022. "The Role of Attention Control in Complex Real-World Tasks." *Psychonomic Bulletin & Review* 29 (4): 1143–1197. https://doi.org/10.3758/s13423-021-02052-2

Esterwood, C., and L. P. Robert. 2021. "Do You Still Trust Me? Human-Robot Trust Repair Strategies." 2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN), 183–188. https://doi.org/10.1109/RO-MAN50785.2021.9515365

Esterwood, C., and L. P. Robert. 2022. "A Literature Review of Trust Repair in HRI." 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), 1641–1646. https://doi.org/10.1109/RO-MAN53752.2022.9900667

Gielo-Perczak, K., and W. Karwowski. 2003. "Ecological Models of Human Performance Based on Affordance, Emotion and Intuition." *Ergonomics* 46 (1-3): 310–326. https://doi.org/10.1080/00140130303536

Kim, P. H., D. L. Ferrin, C. D. Cooper, and K. T. Dirks. 2004. "Removing the Shadow of Suspicion: The Effects of Apology versus Denial for Repairing Competence- versus Integrity-Based Trust Violations." *The Journal of Applied Psychology* 89 (1): 104–118. https://doi.org/10.1037/0021-9010.89.1.104

Kohn, S. C., D. Quinn, R. Pak, E. J. de Visser, and T. H. Shaw. 2018. "Trust Repair Strategies with Self-Driving Vehicles: An Exploratory Study." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 62 (1): 1108–1112. https://doi.org/10.1177/1541931218621254

Lee, J. D., and K. A. See. 2004. "Trust in Automation: Designing for Appropriate Reliance." *Human Factors* 46 (1): 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lyons, J. B., I. A. Hamdan, and T. Q. Vo. 2023. "Explanations and Trust: What Happens to Trust When a Robot Partner Does Something Unexpected?" *Computers in Human Behavior* 138: 107473. https://doi.org/10.1016/j.chb.2022.107473

Merritt, S. M., A. Ako-Brew, W. J. Bryant, A. Staley, M. McKenna, A. Leone, and L. Shirase. 2019. "Automation-Induced Complacency Potential: Development and Validation of a New Scale." *Frontiers in Psychology* 10: 225. https://doi.org/10.3389/fpsyg.2019.00225

Nayyar, M., and A. R. Wagner. 2018. "When Should a Robot Apologize? Understanding How Timing Affects Human-Robot Trust Repair." In: Ge, S., *et al.* Social Robotics. ICSR 2018. Lecture Notes in Computer Science(), vol 11357. Springer, Cham. https://doi.org/10.1007/978-3-030-05204-1_26

Nomura, T., T. Kanda, T. Suzuki, and K. Kato. 2008. "Prediction of Human Behavior in Human–Robot Interaction Using Psychological Scales for Anxiety and Negative Attitudes toward Robots." *IEEE Transactions on Robotics* 24 (2): 442–451. https://doi.org/10.1109/TRO.2007.914004

Norman, D. A., and D. G. Bobrow. 1975. "On Data-Limited and Resource-Limited Processes." In *Cognitive Psychology* 7 (1): 44–64. https://doi.org/10.1016/0010-0285(75)90004-3

O'Neill, T., N. McNeese, A. Barron, and B. Schelble. 2022. "Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature." *Human Factors* 64 (5): 904–938. https://doi.org/10.1177/0018720820960865

Pak, R., N. Fink, M. Price, B. Bass, and L. Sturre. 2012. "Decision Support Aids with Anthropomorphic Characteristics Influence Trust and Performance in Younger and Older Adults." *Ergonomics* 55 (9): 1059–1072. https://doi.org/10.1080/00140139.2012.691554.

Petty, R. E., and J. T. Cacioppo. 1986. "The Elaboration Likelihood Model of Persuasion." In *Advances in Experimental Social Psychology*, edited by L. Berkowitz, Vol. 19, 123–205. Academic Press. https://doi.org/10.1016/S0065-2601(08)60214-2

Robinette, P., A. M. Howard, and A. R. Wagner. 2015. "Timing is Key for Robot Trust Repair." *Social Robotics* : 574–583. https://doi.org/10.1007/978-3-319-25554-5_57

Ryan, R. M., and E. L. Deci. 2000. "Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions." *Contemporary Educational Psychology* 25 (1): 54–67. https://doi.org/10.1006/ceps.1999.1020

Schelble, B. G., J. Lopez, C. Textor, R. Zhang, N. J. McNeese, R. Pak, and G. Freeman. 2022. "Towards Ethical AI: Empirically Investigating Dimensions of AI Ethics, Trust Repair, and Performance in Human-AITeaming." *Human Factors*: 187208221116952. https://doi.org/10.1177/00187208221116952

Singh, I. L., R. Molloy, and R. Parasuraman. 1993. "Automation- Induced "Complacency": Development of the Complacency-Potential Rating Scale." *The International Journal of Aviation Psychology* 3 (2): 111–122. https://doi.org/10.1207/s15327108ijap0302_2

Slamecka, N. J., and P. Graf. 1978. "The Generation Effect: Delineation of a Phenomenon." *Journal of Experimental Psychology: Human Learning and Memory* 4 (6): 592–604. https://doi.org/10.1037/0278-7393.4.6.592

Skraaning, G., and G. A. Jamieson. 2021. "Human Performance Benefits of the Automation Transparency Design Principle: Validation and Variation." *Human Factors* 63 (3): 379–401. https://doi.org/10.1177/0018720819887252

Xu, J., and A. Howard. 2022. "Evaluating the Impact of Emotional Apology on Human-Robot Trust." 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), 1655–1661. https://doi.org/10.1109/RO-MAN53752.2022.9900518

Zhang, Z., R. Tian, and V. G. Duffy. 2022. "Trust in Automated Vehicle: A Meta-Analysis." In *Human-Automation Interaction: Transportation*, edited by V. G. Duffy, S. J. Landry, J. D. Lee, and N. Stanton, 221–234. Cham: Springer International Publishing.