NORTHWESTERN UNIVERSITY

“Alexa, how can I trust you again?”

Trust Repair in Human-AI Teams

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Media, Technology, and Society

By

Alexa Marie Harris

EVANSTON, ILLINOIS

September, 2023

Abstract

The advent of advanced computing and AI has led to social technologies becoming agentic teammates in human-autonomy teams. Interpersonal trust, vital for team functioning, is crucial in determining these teams' success or failure. Trust, while essential, can be easily broken and requires maintenance and repair. This dissertation addresses two questions: Which factors drive trust reparation? And, how can AI teammates effectively navigate trust reparation? An integrative review of trust literature is presented, providing a framework for understanding human-autonomy team trust reparation. Hypotheses are developed and tested in a laboratory experiment consisting of two studies.

The first study employs an MTurk sample and a vignette study to fine-tune manipulations of trust violations and reparative responses. The second study uses Wizard of Oz methodology with a live team of participants and a confederate AI. The findings contribute to understanding the complex interplay between response behavior, violation type, and attributions in AI trust violations. The findings from Study 1 suggest that team members' attributions of stability and controllability to an AI's behavior in response to a trust violation depend on the type of response given by the AI and the type of violation committed. And the findings for Study 2 are inconclusive due to small sample size.

In summary, this dissertation establishes the foundation for future research on trust violations in human-autonomy teams, providing guiding principles for trust reparation behavior for AI.

Acknowledgements

First, I'm so thankful to my academic advisers who led me through this dissertation and the journey that it has truly been. Thank you to my very exceptionally brilliant adviser, Dr. Leslie DeChurch, who taught me how to conduct myself as an academic. Thank you to Dr. Noshir Contractor for providing me with guidance and opportunity. And thank you to Dr. Nina Lauharatanahirun for being both a friend and an adviser in my dissertation journey.

I also want to thank the team that made my dissertation possible. A special thank you to Carli Kelley, Ximena Munoz, and Fumika Hoshi, among many other undergraduate researchers (Ariel Gordon, Charlotte Cagliostro, Cindy Wu, Eva Offutt, Kim Nguyen, Louise Oh, Mackenzie Matheson, Ruby Chao,Seora Kim, Tano Barendsen-Rossi, Shannon Lackey, and Wilson Ting - among many others over the years) without whom this large project would not be accomplished.

I want to thank my friends, from those who supported me from the very beginning, to those who I met along the way. It has been one hell of a journey. Thank you for your contagious brilliance, camaraderie, care, and support.

Finally, I want to thank my mother and my brother. To my loving mom, thank you for always believing in me. For never giving up on me during the not-so-easy times. And for always allowing me a fresh start. You are an amazing person and I will forever be grateful for your love, care, and guidance in life. And to my brother, thank you for always providing me with perspective, love, and the life advice only an older brother can give.

This dissertation is based on work supported by the Army Research Laboratory (W911NF1920140). The views, opinions, and/or findings contained in this report are those of the authors and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documents.

Table of Contents

List of In-Text Tables and Figures

Appendices

Executive Summary

Two studies aimed to examine team members' attributions of stability and controllability to an AI's behavior following trust violations, focusing on the effects of response behavior (denial vs. apology) and violation type (competence vs. integrity). The findings from study 1 suggest that stability attributions are higher when an AI denies wrongdoing, while competence violations lead to lower stability and higher controllability attributions compared to integrity violations. However, the study did not find significant differences in controllability attributions between denial and apology conditions, nor were most hypotheses regarding the relationships between attributions and expectancy for change or sympathy supported. Findings from study 2 are inconclusive and require a higher sample size.

The study highlights the need for further research to explore factors influencing attributions in AI trust violations and understand the relationships between attributions, expectancy for change, and sympathy. From a practical standpoint, these findings can inform the development of AI systems, trust repair strategies, and better management of human-AI team dynamics.

Introduction

We are on the cusp of a new genre of teams that combines the potential of people and intelligent machines. Advances in artificial intelligence (AI) and robotics are paving the way for socially adept agents to collaborate and join teams (Bohannon, Fitzhugh, & DeCostanza, 2019; Breazeal, 2004; Rahwan et al., 2019). Humans will interact with, rely on, and trust artificially intelligent, machine teammates in ways that are presently only imaginable by watching or reading science fiction. Greater AI integration has implications for all aspects of society, particularly the workplace, where teams have become the basic unit of work in most organizations (Mathieu, Hollenbeck, van Knippenberg, & Ilgen, 2017).

A team is defined as "an intact social system, complete with boundaries, interdependence for some shared purpose, and differentiated member roles (Hackman, 2012, p. 429)." DeCostanza and colleagues (2018) define human-autonomy teaming to be, "[...] composed of human team members as well as distributed sensors, robots, unmanned aerial vehicles (UAVs), autonomous vehicles, intelligent assistants, and other advanced technologies that can perform taskwork as part of the larger team, while we reserve the term technology for those devices, software, protocols, and other interventions that target the members of the team with the goal of improving team processes. It is entirely possible that a technology will also be a team member [...] (p. 3)."

A central characteristic of team or human-autonomy team success is the notion of trust, defined as, "[Trust is] the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other part (Mayer, Schoorman, & Davis,

1995, p. 217).” Trust has been shown to be an integral part of effective team functioning (Mach, Dolan, & Tzafrir, 2010) and has been shown to mediate important team processes and outcomes such as cohesion and performance (Dirks & Ferrin, 2001). It is also an intuitive construct; if you don’t trust your teammate, how will you work with them effectively? Thus, when trust is lowered from a previously heightened state, team members should seek to repair trust. However, the process of trust repair typically requires evidence that the person responsible for lowering trust is trustworthy, and this is often difficult to establish (Medina, 2020).

While team dynamics were already complicated when trying to understand a team full of people, teams are perhaps about to become even more complicated with the incorporation of artificially intelligent teammates. Table 1 details the different types of trusting relationships possible in a mixed initiative team, or a team that is composed of humans and AI of various team sizes; note that the trustor in all of these relationships is a person, though, in the future research may develop models of trust from the AI’s perspective. These AI teammates will function just as any other member of the team would, and, therefore, it is not a matter of *if* the AI teammate breaks team member trust, but matters of *when* and *how*. For example, a violation of trust in a human-AI team can happen simply because expectations for behavior are left unmet due to unaligned goals; as AI on these teams will be fully autonomous, their goals may be in conflict with their human teammate’s goals. In this dissertation, I seek to find the behavioral responses an AI can enact to restore trust once it has broken it.

Table 1

Trusting Relationships Possible in a Mixed Initiative Team

Interaction Type	Level	Trustor(s)	Trustee(s)
Interpersonal	Dyadic	Human	A person that will be or is the cause of an event.
Human-AI Trust	Dyadic	Human	An autonomous artificial intelligence in any

			form of embodiment that will be the cause of an event.
Team Trust	Team	Human	Other members of the team as trustees lead to the emergence of the team as a whole becoming the trustee.
Trust in Human-Autonomy Teaming	Team	Human	Other members of the team, including a dyadic relationship between the trustor and an AI teammate trustee, lead to the emergence of the team as a whole becoming the trustee.

To ground this work in an example, albeit a futuristic example, of a human-autonomy team consider the movie *Interstellar* (Nolan, C. (2014). *Interstellar*. Paramount Pictures.), where the Earth is slowly being rendered uninhabitable and humanity must send its best and brightest on a mission to find a new home. They send Cooper, an accomplished NASA pilot, a team of researchers, and two AI crew members named TARS and CASE aboard a ship named the Endurance on an exploratory mission. After an unforeseen problem, the crew must work together to survive and return home to Earth. In one specific example of teamwork, after the Endurance begins spinning out of control and falls out of orbit, TARS - who is on the Endurance, CASE, and the rest of the crew work together to dock their smaller ship to the Endurance. Below is the transcript beginning with CASE's response to Cooper brazenly accelerating toward the spinning Endurance.

> CASE: Cooper there's no point in using our fuel -
> Cooper: [CASE] Analyse the endurance's spin.
> Copilot: Cooper, what are you doing?
> Cooper: Docking...
> CASE: The endurance rotation is 67- 68 rpm.
> Cooper: Ok, get ready to match our spin with the retro thrusters.
> CASE: It's not possible.
> Cooper: No, it's necessary.
> TARS: The Endurance is hitting the stratosphere.
> Copilot: She's got no heat shield!
> Cooper: CASE, you ready?

CASE: Ready!
CASE: Cooper! This is no time for caution!
Cooper: CASE if I black out, you take the stick. TARS get ready to engage the docking mechanism.
CASE: The Endurance is starting to heat.
Cooper: 20 feet out!
TARS: I need 3 degrees starboard, Cooper.
Cooper: 10 feet out!
TARS: Cooper, we are.... lined up!
Cooper: Initiate spin!
Cooper (on the brink of passing out from the g force): Come on TARS! Come on TARS! Come on! Easy now! Easy..... [to CASE] Retro thrusters! Main engines on! Pushing out of orbit! Come on baby... Killing main engines! Ok, we're out of orbit!

The transcript from *Interstellar* exemplifies a scenario in which the crew has a shared task at hand and must use their shared understanding about equipment, task goals, and task subgoals to work together and survive. They have a shared commitment to the task and generally there is a shared belief that if they work together, they will survive. Because this is a life or death circumstance, this is an extreme example of trust. Everyone eventually trusted Cooper's judgment; Cooper, in return, trusted his mixed crew of humans and AI to execute the plan. Without that trust, the plan would have fallen apart and possibly cost them their lives. One can easily imagine this scenario going differently had any of the crew members broken Cooper's trust shortly before he made his decision.

Trust

The crew in the Interstellar example can be conceptualized as a team. A team can be understood as a "complex dynamic systems that exist in a context, develop as members interact over time, and evolve and adapt as situational demands unfold." (Kozlowski & Ilgen, 2006, p. 78). For a team to function optimally, there must be an establishment of trust as demonstrated in the above example. Trust, as a concept, is easily understood; it is intuitive and often occurs without much consideration in everyday life. Among the most common definitions of trust is

"the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other part (Mayer, Schoorman, & Davis, 1995, p. 217)." Table 2 presents some of the prominent definitions of trust that have appeared in, or been cited within, research on traditional teams and human-ai teams.

Table 2
Exemplar Definitions of Trust

Citation	Trust Type	Definition
Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709-734.	Interpersonal Trust	"[Trust is] the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other part (p. 217)."
McAllister, D. J. (1995). Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal*, 38(1), 24-59.	Interpersonal Trust	"Interpersonal trust is the extent to which a person is confident in, and willing to act on the basis of, the words, actions, and decisions of another. (p. 25)"
Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50-80.	Human-Machine Trust	"A simple definition of trust [...] is the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability. (p. 54)"
Fulmer, C. A., & Gelfand, M. J. (2012). At what level (and in whom) we trust: Trust across multiple organizational levels. *Journal of Management*, *38*(4), 1167-1230.	Team Trust	"[Team trust is] a shared psychological state among team members comprising willingness to accept vulnerability based on positive expectations of a specific other or others. (p. 1174)"

The academic consensus on interpersonal trust is that there is no unification or agreement on a specific or operational definition of trust, much less a definition of trust that can be applied to all instances of teaming (e.g. a team of all humans, a mixed initiative team) (Colquitt, Scott, & LePine, 2007; Schaefer et al., 2016; Lewis, Sycara, & Walker, 2018). For example, Colquitt, Scott, and LePine (2007) say, "[The] multidisciplinary perspective has created a breadth that strengthens the trust literature, it also has created confusion about the definition and conceptualization of the trust construct, p. 909," attributing the confusion to the interdisciplinary approach to understanding trust. Lewis, Sycara, and Walker (2018) hold a position that, "[Trust] appears to hold many components that we may never converge on a single, precise, and concise definition of the concept, p. 5," and that adding in a machine trustee makes this relationship even more complex to the point we may not ever get to a convergence and may simply require multiple definitions and frameworks depending on the situation. Lewis and colleagues go as far as saying, "In the ultimate form of trusted autonomous systems, the parties of a trusting relationship are both autonomous; thus, both parties need to establish trust in themselves, and then in each other. If one party is a human and the other is a machine, the machine needs to trust the human (machine-human trust) and the human needs to trust the machine (human-machine trust). Therefore, to merely assume that the machine needs to respect what trust is in a human system limits our grasp on the complexity of trust in trusted autonomy (Lewis, Sycara, & Walker, 2018, p. 6)," implying that research should focus on the direction of the trusting relationship and develop two different frameworks rather than a generalizable model of trust. This problem space is incredibly complex, and to make any progress on understanding the trust process, researchers must break off smaller pieces.

As is apparent from the array of definitions, trust is a difficult construct to define. Some argue that there has yet to be one concrete and agreed upon definition that captures all of the aspects (Lewis, Sycara, & Walker, 2018; Colquitt, Scott, & LePine, 2007). One reason for this may be because trust between two people differs from trust in technology depending on which form of embodiment (Schaefer et al., 2016; Glikson & Woolley, 2020), or that environments and organizational settings can dictate antecedents of interpersonal trust (Meyerson, Weick, & Kramer, 1996) and when looking across fields they appear to be inconsistent. The trust literature is scattered across many decades, many fields, and many articles offering their own definitions of trust. To illustrate this variance, a prominent definition of trust from the human factors discipline is, "[...] the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability. (Lee & See, 2004, p. 54)"; from psychology, "[Trust is a] state of perceived vulnerability or risk that is derived from an individual's uncertainty regarding the motives, intentions, and prospective actions of others on whom they depend (Kramer, 1999, p. 571); or from computer science, "a strong human belief in the reliability, truth, or ability of an autonomous system (Shahrdar et al., 2018, p. 2). These definitions exemplify the variance in complexity, what is included, and what is excluded from the conception of trust.

Rather than review all of the various definitions offered across many decades, journals, and fields, I utilize a prominent conceptual definition of trust, "[Trust is] the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party (Mayer, Schoorman, & Davis, 1995, p. 217)," and a theoretical framework which bifurcates trust into two main components, cognitive and affective (McAllister,

1995). While the trust literature includes many splintering definitions and lines of research, this definition and framework have the largest and most robust body of literature to support them. Thus, suffice it to say, there has yet to be complete consensus on what exactly trust is; however, for this dissertation, this will be the framework and definition used.

Human-AI Trust

To begin to understand what human-AI trust entails, we first must define artificial intelligence (AI). AI is the overarching general intelligence of a machine with four advanced capabilities: 1) interacting with the surrounding environment and gathering information; 2) interpreting this information, recognizing patterns, inducing rules, or predicting events; 3) generating results, answering questions; or giving instructions to other systems; and 4) evaluating the results of their actions and improving the machine's decision systems to achieve specific objectives (Ferrás-Hernández, 2018). Due to these four component capabilities, the decision making algorithms are dynamic and react situationally to stimuli in the environment. It has been argued that the ability to learn and react dynamically to the environment makes AI a social actor in that environment (Rahwan et al., 2019). As with any actor interacting with any other actor in any given environment, there must be a trusting relationship between actors for seamless collaboration to occur.

A major difference between interpersonal trust and human-AI trust is that AI can present itself in many different forms or embodiments. As the embodiment of the AI dictates capabilities and interaction modalities, embodiment must also be defined and considered at each level of embodiment. Traditional perspectives on embodiment defined it as, "a term used to refer to the fact that intelligence cannot merely exist in the form of an abstract algorithm but requires a physical instantiation, a body (Scheier, & Pfeifer, 1999, p. 649)." This instantiation is not

physical in the sense that it is tangible in the physical world, but may also exist in a simulated physical body in the virtual world (Ferber, & Weiss, 1999). However, as seen currently, there are completely disembodied algorithms that interact with people constantly and shape real world decisions (e.g. Facebook, Google, or even IBM's Watson). Embodiment, then, can be split into three categories: 1) robotic, automation or robotic AI that occupies space in the physical world and may manipulate the physical environment around it given the physical appendages to do so; 2) virtual embodiment, a virtual agent existing and interacting solely with the virtual world; or 3) an invisible embedded instantiation that interacts with the virtual world outside of visual representation (Glikson & Woolley, 2020).

Early work on embodied computer systems focused on either human-robot trust or human-automation trust. Automation is the predecessor of AI in that it is embodied and functions without direct human intervention; however, AI is distinct from automation given the dynamism of its decision making. Formally defined automation is "technology that actively selects data, transforms information, makes decisions, or controls processes" (Lee & See, 2004, p. 50). The definitions for AI and automation may seem markedly similar, the distinction lies in the ability to make dynamic decisions rather than selecting from preprogrammed options. For example, robotic AI may look identical to automated robotics in form, however, AI controlled robotics would be able to gather information from its surrounding environment, analyze it and recognize patterns, perform actions based on those analyses, and then evaluate the efficacy of those actions and adapt its behavior accordingly. In contrast, an automated robot in the same environment would simply stick to whichever action it was preprogrammed to do that best achieves the desired outcome in the given situation. Because of these differences, and differences in embodiment, the antecedents for automation trust are different than that of human-AI trust.

Meta analytic findings on human-automation trust, which includes intelligent automation that can be considered AI, indicate a three-factor model composed of human factors, partner factors, and environmental factors (Schaefer et al., 2016). Each factor is a broad category of antecedents containing specific sub-factors. The “human factor” is composed of the human trustor’s traits, states, cognitive factors, and emotive factors that are discussed in the interpersonal trust section of this paper. Each of these human related factors may vary from person to person. The “partner factor" is composed of all of the factors attributed to the partner that function as antecedents to trust; the “partner” is the automation or technological trustee in any form of embodiment. Partner related antecedents are the features (e.g. mode of communication, appearance, intelligence, personality, and level of automation) and the capabilities of the technological trustee (e.g. behavior, reliability, and feedback). The final component, the “environmental factor,” is composed of antecedents attributed to the team or the organizational environment in which the trust relationship is occuring (e.g. role interdependence, team composition, mental models, culture, and group membership). Schaefer and colleagues note that if the automation is functioning within a team, the team becomes context that influences the human-automation trust, although there were insufficient empirical articles in this area to properly meta-analyze.

Advances in AI are pushing AI away from its automation predecessor. However, AI that parallels human intelligence in any given situation (strong AI) has not been developed yet, fragments of human intelligence level AI have (weak AI) (Raj & Seamans, 2019). Weak AI functions primarily on machine learning algorithms, algorithms that take in data and adjust accordingly. Machine learning as a technique typically requires large amounts of labeled data

from a specific situation to train the algorithm on. The algorithm then uses the patterns from the training data to assess new data in a similar situation (Brynjolfsson & Mitchell, 2017).

Human-AI trusting relationships should function similarly to interpersonal, human-human trusting relationships and will only close the gap between the two as AI becomes more and more intelligent.

Trust Violation and Repair

Interpersonal trusting relationships are tenuous and easily broken, even in very simple everyday interactions (McKnight, 1998; Kim, Ferrin, Cooper, & Dirks, 2004). For trust to exist, both parties must be interdependent, there must be an element of risk involved, and there must be a desired outcome such that if one party neglects their duties, or fails in their obligations entirely, the other party loses something of value (Rousseau et al., 1998). Therefore, one can conceptualize trust in phases. The first phase is the establishment (Mayer et al., 1995; Lewicki, Tomlinson, & Gillespie, 2006), then a maintenance period in which the trustee must perform certain actions to maintain the trustor's trust (Simpson, 2007). This maintenance phase is cyclical and repeats until trust has been violated. It is this space of trust violation and loss in which this dissertation is settled. When trust has been broken, a party has lost something of value due to the fault of another party that they had entrusted to successfully carry out their duties and must decide two things: 1) whether to repair trust or abolish the relationship entirely; and 2) if the trustor decides to repair trust, how much trust do they restore.

Scholars believe there are three types of trust violations: 1) competence, 2) integrity, and 3) benevolence-based trust violations (Ozturk & Ozmen 2013). These three categories of trust violations are congruent with Mayer, Schoorman, and Davis' (1995) framework. To date, the majority of trust research, especially empirical research, has focused on competence-based

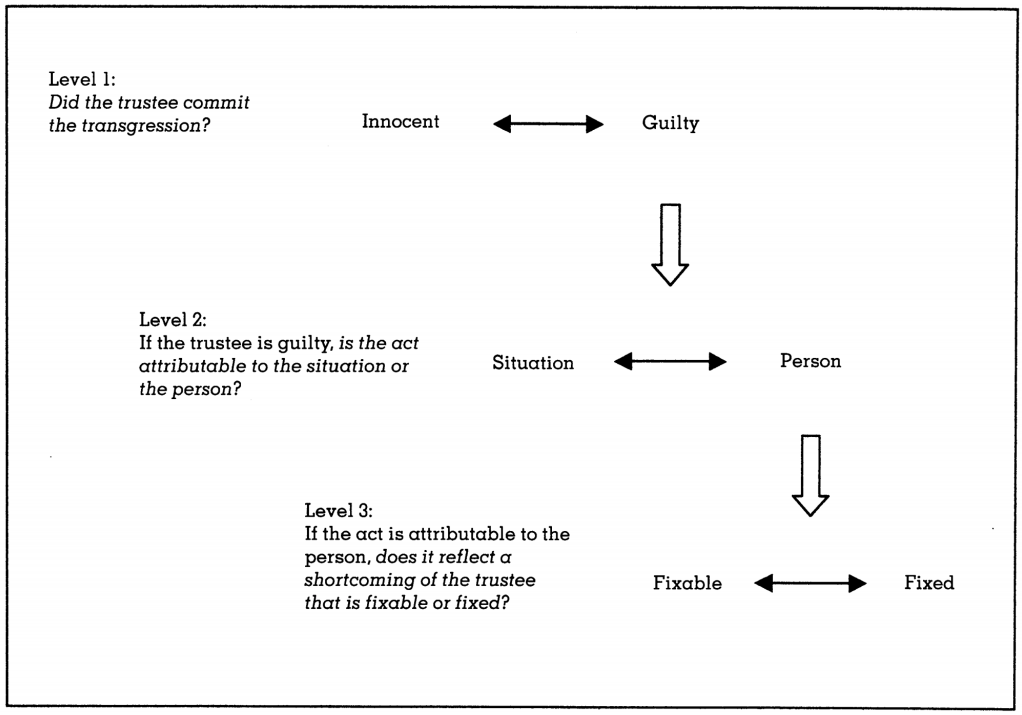
violations (Colquitt et al., 2007); the remaining two violation types remain underdeveloped. Trust repair for the purposes of this dissertation is defined as the process in which an individual restores trust in another individual that has not met their expectations.

Trust reparation, after a trust violation, can be understood as the process of reestablishing trust. At first glance, this process may seem similar to that of the first time trust is established; however, there are many new factors that are involved that can affect the decision to repair trust or abolish the trusting relationship entirely. Much of the extant research focuses on the first part of the trusting relationship, where trust is initially established. The small portion of literature that focuses on the process of trust repair either focuses on the perceived trustworthiness of the trustee's behavior, such as the efficacy of an apology, or the attributions cast upon the trustee by the trustor. Thus there are three streams of research on the process of trust repair: 1) the behavior of the trustee, 2) the attribution of blame and perception of the situation by the trustor, and 3) the individual differences of the trustor that make them more or less likely to trust.

Research on the process of trust repair shows a contingency table which depicts the possible outcomes of a negotiation between the trustee and the trustor and their willingness to repair the trusting relationship, seen in Figure 2 (Kim, Dirks, & Cooper, 2009). This contingency is reliant on the bilateral model of trust repair, Figure 3, composed of three levels. The first being whether or not the trustee committed the transgression, are they guilty or innocent. The second level asks the question, if the trustee is guilty, is the act attributable to the trustee or to the situation? And finally, the third level is, if the act is attributable to the person and not the situation, is the act fixable or fixed? This bilateral model of trust repair utilizes similar frameworks as Jackson & LePine's model based on attribution theory. The trustor is essentially processing through the information and attributing the outcome to either the person or the

situation.

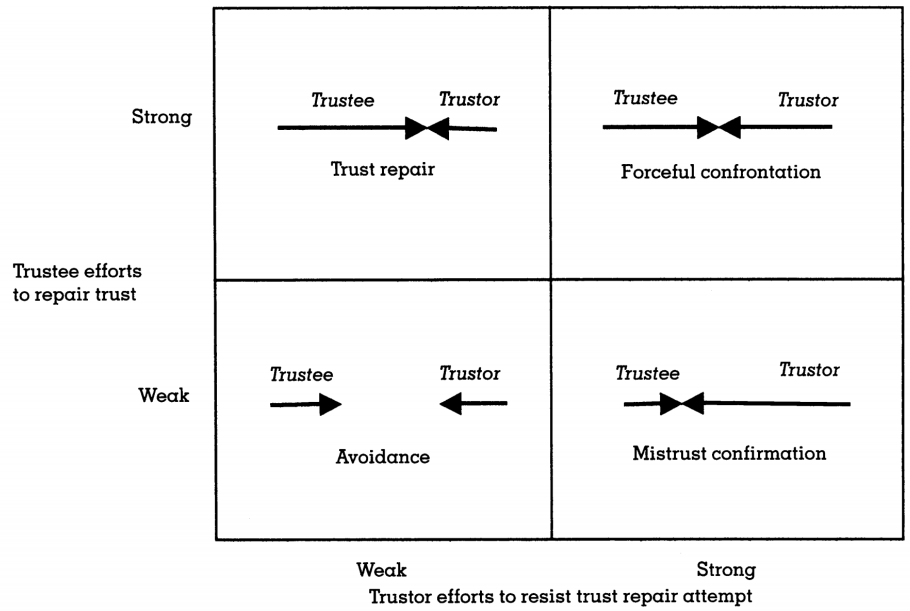
Figure 1
Bilateral Model of Trust Repair



Note: Model from Kim, Dirks, and Cooper's (2009) paper on trust repair.

Figure 2

Potential Outcomes of Negotiation Efforts by a Trustor and Trustee



Note: Contingency table from Kim, Dirks, and Cooper's (2009) paper on trust repair.

Similar to Kim, Dirks, and Cooper's (2009) contingency table based on the willingness of both parties to repair trust, there are a set number of behavioral options that a trustee can perform after a trust violation; this is the same for an AI teammate or a human teammate, see Table 2 (De Visser, Pak, & Shaw, 2018). Some of these behaviors, such as apologizing, can act as a strong indicator to the trustor that they should repair trust. Other behaviors, like downplaying the transgression can lead to signals that the trustee has no interest in repairing the trust, and ultimately signaling untrustworthy behavior.

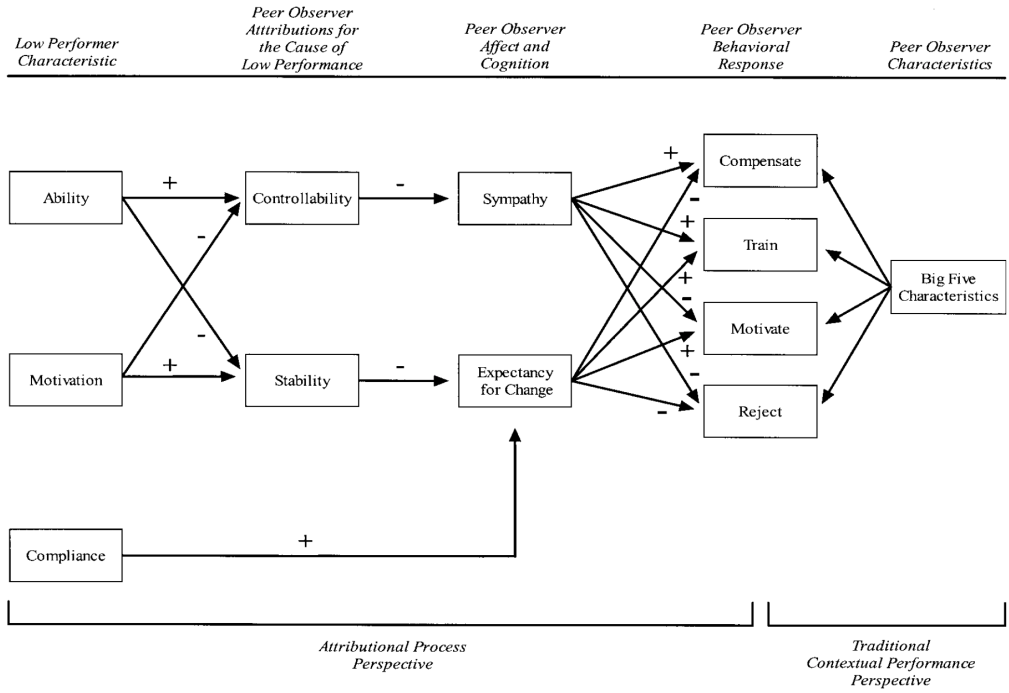
According to a taxonomy of behavioral outcomes following a breach in trust in an organizational setting there are four possible behaviors one can expect from the trustor: compensate, train, motivate and reject (Jackson & LePine, 2003). The actual decision that must be made following a violation in trust is binary; either the trustor decides to trust the trustee again or decides to terminate the trusting relationship entirely. In an organizational setting such

as a team tasked to accomplish a specific goal, once the decision has been made to trust again, the trustor may not restore trust to prior levels. This could result in behaviors such as compensating, retraining the trustee, or attempting to motivate the trustee.

Jackson and LePine (2003) utilize attribution theory (Weiner, 1972) as a framework in which trustors decide how to handle a poor performer. This poor performance functions as a violation of trust and thus functions as a model of trust repair as well. In an organizational setting, the attributions made will be based on the trustee's perceived level of ability, motivation, how much control the trustee was perceived to have in the situation, how stable the behavior of the trustee has been, the trustor's sympathy for the trustee, and the trustor's expectation for the trustee to change; see Figure 1. These attributional judgements will all affect how much trust is restored in the trustee post trust violation.

Figure 3

Attributional Model of Peer Poor Performance



Note: Model from Jackson and LePine's (2003) paper on perceptions of peer poor performance.

At this point it is pertinent to parse out what exactly is meant by trust repair as forgiveness, trait forgiveness, and reconciliation all deal with similar yet different processes than the trust repair detailed in this dissertation. Forgiveness and trait forgiveness (Enright, Freedman, & Rique, 1998) is concerned with individual differences that predict the likelihood of absolution after a violation. That is, an unconditional, volitional response to another person after that person has violated trust. The unconditional nature of forgiveness has led scholars to believe that forgiveness resides wholly within the trustor and only influenced by the trustor's aversion to negative feelings associated with the violation in trust. That is, the trustor doesn't want to feel upset about what happened and chooses to just 'let it go' in an attempt to reduce negative feelings.

Reconciliation is more related to trust repair than forgiveness as reconciliation details the process of negotiating trust repair between the trustor and the trustee. That is, reconciliation is a bilateral process that requires goodwill by both the trustee and the trustor after a trust violation has occurred (Fincham, 2000). The construct of trust depletion (Hirschman, 1984), is the extent that trust has diminished due to neglected trust relationship maintenance due to an extended period of time apart. Returning to the battery metaphor, trust depletion is the process of the battery being drained slowly over an extended period of time left sitting. If the trustor and the trustee has not been in contact with the trustee in a long period of time, trust will slowly decay.

Finally, trust erosion is the amount of trust lost after a trust violation, grounded in the trustor's knowledge of the trustee's prior performance (Elangovan, Auer-Rizzi & Szabo, 2006). Trust erosion assumes a state of high initial trust, and, after a trust violation, it is the amount of trust lost. This construct may be the most related to the constructs developed in this dissertation; however, trust erosion is the imperfect inversion of trust repair. It stands to reason that the factors

that would influence the amount of trust lost do not perfectly mirror the factors that repair the trust lost.

Table 3

New and Related Constructs

Author(s)	Construct (State/Trait)	Definition
Mayer, Davis, & Schoorman, 1995	Trustworthiness	Attributes of the trustee perceived by the trustor. Composed of: ability, integrity, and benevolence.
Gills, Bois, Finegan, & McNally	Propensity to Trust	The trustor's unique likelihood of trusting a trustee.
Hanoch, Johnson, & Wilke, 2006	Propensity for Risk Taking Behavior (PRTB)	A combination of traits in the trustor that moderate the likelihood of accepting risk.
Tomlinson & Mayer, 2009	Trust Repair	The period of time and process after the trustor's trust has been violated where the trustor must decide to trust or not to trust the trustee again.
Mayer, Davis, & Schoorman, 1995	Trust	"[Trust is] the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other part (p. 217)."
--	Trust Establishment	The front end of the trust cycle. Trust establishment is moderated by propensity to trust and the perceived trustworthiness of the trustee.
Enright, Freedman, & Rique, 1998	Forgiveness	An unconditional, volitional response to another person.

Elangovan, Auer-Rizzi & Szabo, 2006	Trust Erosion	The factors that influence the amount of trust lost after a trust violation, grounded in the trustor's knowledge of the trustee's prior performance. Trust erosion assumes a state of high initial trust.
Elangovan & Shapiro, 1998	Trust Violation	Trust is violated when the trustor perceives the trustee as acting in a way that does not fulfill their expectations.
Hirschman, 1984	Trust Depletion	The slow decline of trust due to little or no contact between the trustee and the trustor for an extended period of time.
Fincham, 2000	Reconciliation	Reconciliation is a bilateral process that requires goodwill by both the trustee and the trustor after a trust violation has occurred.

Trust in the Context of a Team and Team Level Trust

Trust between two people does not occur in a void; however, this point is especially pertinent when considering interpersonal trust in a team context. When interpersonal trust occurs in a team context, the relationship forms with information about other members of the team's trust relationships as well; trust in the presence of other interpersonal trust relationships, put simply (Gupta, Ho, Pollack, & Lai, 2016). For example, Bliese (2000) proposes that attitudes among team members are non‐independent, such that one member's trust in the team is expected to affect and be affected by other members' trust in the team. This can be conceptualized as the team as context for interpersonal trust; or that trusting relationships are influenced by the presence of adjacent trusting relationships. But the team is far more important than just simply context.

Theories of emergent team properties take a Gestaltian perspective toward teams, that the whole is greater than the sum of its parts (Koffka, 2013), as well as a multi-level approach (Kozlowski & Klein, 2000; Costa, Fulmer, & Anderson, 2018). Trust, for example, can exist at multiple levels within a team, dyadic, one team member's trust in the team as a whole, and team-level trust. At the dyadic level, trust can be established and held between two individuals within the team (e.g. "I trust Danica more than I trust Thomas even though both are members of my team."). An individual may also hold a feeling of overall trust in their team (e.g. "If I struggle on this task, I trust my team will help me."). Team-level trust can be described as the team's collective level of trust in the team as a whole, measured by averaging each team member's trust in the team, or by averaging the trust across dyads, measured by averaging the trust from each dyadic trust relationship (Feitosa et al., 2020; Carter, Carter, & DeChurch, 2018). Therefore, team-level trust emerges as a distinct level of trust from both of these types of trust relationships.

Trust at the team-level, team trust, refers to "a shared psychological state among team members comprising willingness to accept vulnerability based on positive expectations of a specific other or others (Fulmer & Gelfand, 2012, p. 1174)". It has been formally defined as, "Team trust is defined as the shared willingness of the team members to be vulnerable to the actions of the other team members based on the shared expectation that the other team members will perform particular actions that are important to the team, irrespective of the ability to monitor or control the other team members (Bruer, Huffmeier, Hibben, & Hertel, 2016, p. 7)." based on Mayer, Schoorman, and Davis's interpersonal trust definition at the dyadic level. Team trust emerges from continued interactions with members of the team (Williams, 2001), and is an emergent state because it is a shared construct between all members of the team at some level (Kiffin‐Petersen, 2004; De Jong & Elfring, 2010; Burke et al., 2007; Kozlowski & Klein, 2000).

While team trust is distinct from interpersonal trust, cross-level effects between the two create a loop of effect where team trust influences interpersonal trust and vice versa (Kramer & Kramer, 2010; Costa et al., 2018). In teams that are a part of a broader organization, organizational level changes and organizational structures (e.g. hierarchies) can also influence team and individual trust (Ilgen, 1999; Hardin & Offe, 1999). The direction of the cross-level effect can be conceptualized as being either "bottom up" or "top down" supporting the appropriate application of a multi-level perspective (Chen & Kanfer, 2006; Chen & Tesluk, 2012). This perspective, however, makes untangling antecedents of team trust exceptionally difficult.

The emergent nature of team trust dictates that antecedents for dyadic trust are largely the same for team trust, thus influencing team trust from the bottom up; but interpersonal trust antecedents can also be aggregated to the team-level and directly influence team trust (Costa et al., 2018). Propensity to trust, trustworthiness of team members, and the history of the dyadic trust relationships within the team are all individual or dyadic antecedents of team-level trust. Team-level antecedents of team trust are split into two categories, social and structural (Costa et al., 2018). Social antecedents of team trust consist of team processes (Marks, Mathieu, & Zaccaro, 2001), team climate (Hülsheger et al., 2009; Edmondson, 1999), and leadership practices (Dirks & Ferrin, 2002; Gillespie & Mann, 2004); and structural antecedents of team trust consist largely of team composition (Newell et al., 2008) and the medium in which the team communicates (e.g. virtual teams; Wilson et al., 2006).

Due to team trust's multi-level nature, a multi-level model of team trust is necessary to capture the interplay between interpersonal trust at the dyadic level and team trust at the team level (Costa et al., 2018). For example, Costa and colleagues (2018) developed a multi-level model of team trust in which team trust affects both the interpersonal level trust outcomes as well

as team level trust outcomes. Cross level effects such as this also help to illustrate the feedback loop they create. The interpersonal trust then influences team trust via emergence.

Trust Repair at the Team Level in Human-AI Teams

An important transition is happening in the area of technology and teams. This is a shift from viewing technology as a tool to viewing technology as a teammate (Fiore & Whittshire, 2016; Larson & DeChurch, 2020). This may seem like a relatively small shift in thinking, but it will change the way we team; it will also require an even more complex understanding of trust, e.g. what it means when the team dyadically trusts their AI teammate, how the presence of trusting relationships affects adjacent trusting relationships, and how the AI teammate learns to trust its team and its dyadic trusting relationships within the team.

The existing HRI frameworks work well when technology is perceived as a tool, but they do not include perceptions of technology as a teammate (Phillips et al., 2011). Major differences have been found in the way people interact with physically embodied AI, or robots, when in groups compared to when they are interacting secluded and dyadically, such as groups of people are more likely to interact with the robot, exhibit intergroup bias in their interactions with the robots, pay less attention to the robot, and externalize their mental states (Sebo et al., 2020).

The majority of studies in examining trust in human-autonomy teams examine one person working with one AI teammate, but the presence of others can influence dynamics within teams. Although not directly fitting the criteria of a team for this paper, one study found that, when compared, a team of two people played a game versus a robot were less cooperative with the robot than when they played the game alone versus the robot (Chang et al., 2012). This finding demonstrates that the presence of others alters the dynamics between humans and their robotic peer. To fully understand human-AI team dynamics, researchers must push for larger teams.

Due to the likely possibility that the presence of team members alters dynamics, form factors and the capabilities for social interaction of the AI may play an even more important role in team level trust than in dyadic interactions. A robot's personality and emotion can influence trust as demonstrated in prior sections of this paper. Robot personality characteristics are often conveyed verbally, such as collaborativeness, trustworthiness, and warmth requiring a baseline of being able to communicate verbally (Sebo et al, 2020); but the physical representation has also shown to change perceptions of competence and warmth (Bergmann et al., 2012), and people prefer robot behaviors that convey competence and warmth (Scheunemann et al., 2020).

Trust Repair and Team Trust Repair

Trust repair at the dyadic level functions similarly to the first time one decides to trust another; however, due to the temporality of trust repair, that an event must happen to violate trust for the process of trust repair to happen, different antecedent factors are observed and weighed by the trustor during the trust repair process than in the trust respiration process. When someone is trying to gain the trust of another, they will perform certain acts to convince the person they are trying to gain the trust of that they are worthy of being trusted (Ferrin, Bligh, & Kohles, 2008). These actions change slightly when one is trying to regain the trust that they have lost. For example, after trust has been diminished or lost, the trustee may apologize to the trustor, and, in doing so, accept responsibility for their actions in an attempt to show the trustor that they are worthy of being trusted again, this process functions as a negotiation between the trustee and the trustor until trust is either restored or believed to be irreparable and both parties walk away (Kim, Dirks, & Cooper, 2009). For an additional example, in the initial trust establishment phase, an apology would be out of place, and contextually would not make sense, but in the trust repair phase it makes contextual sense.

De Visser and colleagues (De Visser, Pak, & Shaw, 2018) have worked out all of the possible behavioral responses a trustee can exhibit post trust violation; though their list was specific to human-AI trusting relationships, there is no difference in interpersonal trusting relationships. Some of these include behaviors that do not help the trustee redeem their trustor's trust, and, instead lead the trustor to view them as being even more untrustworthy. For example, a trustee that has just violated the trustor's trust may choose to ignore the fact that they have just done so, or they may downplay the severity of their actions (De Visser et al., 2018). Returning to the behavioral outcomes proposed by Jackson and LePine (2003), I have categorized these four outcomes as destructive, reject and motivate, and constructive, train and compensate. I created these categories to reflect outcomes that are beneficial to an AI rather than to a human. For example, a human would benefit from motivating and training, but an AI would not benefit from motivation as it is data driven. It is entirely plausible that the sympathy mediator in Jackson and LePine's model will drop out entirely when interacting with an AI. Table 2 has the full list of behavioral response possibilities as well as marked which ones are destructive or constructive for an AI, however, some responses will aid an individual in appearing trustworthy and others will only detract. It is worth noting that, to date, there are no meta analyses that report the effect sizes of these antecedents.

In addition to the behavioral antecedents, which reside at the individual level within the trustee, there are antecedents which reside within the trustor. After a trust violation has occurred, the trustor must process through the event and attribute the cause, who is at fault, was it out of their control, etc, also noted in Table 2. These trustor antecedents are also found in Jackson and LePine's attribution model described in a prior section, see Figure 1.

It is important to note that the overall process of repairing trust with an AI seems to function the same way as it does for interpersonal trusting relationships, as a negotiation where the trustor weighs trustee attributes and behaviors. The main difference being the weight, or effect sizes, of the antecedents used by the trustor to make their decision. For example, one meta analysis found that initial trust between a person and a physically embodied AI teammate is mostly dependent upon the ability of the AI (Hancock et al. 2013). That is, if the AI is perceived as not being capable of completing tasks, then it is unlikely to be trusted. Thus, it is probable that the antecedents are weighted differently when the trustee is a human or an AI.

Table 4

Behavioral Responses to Trust Violations

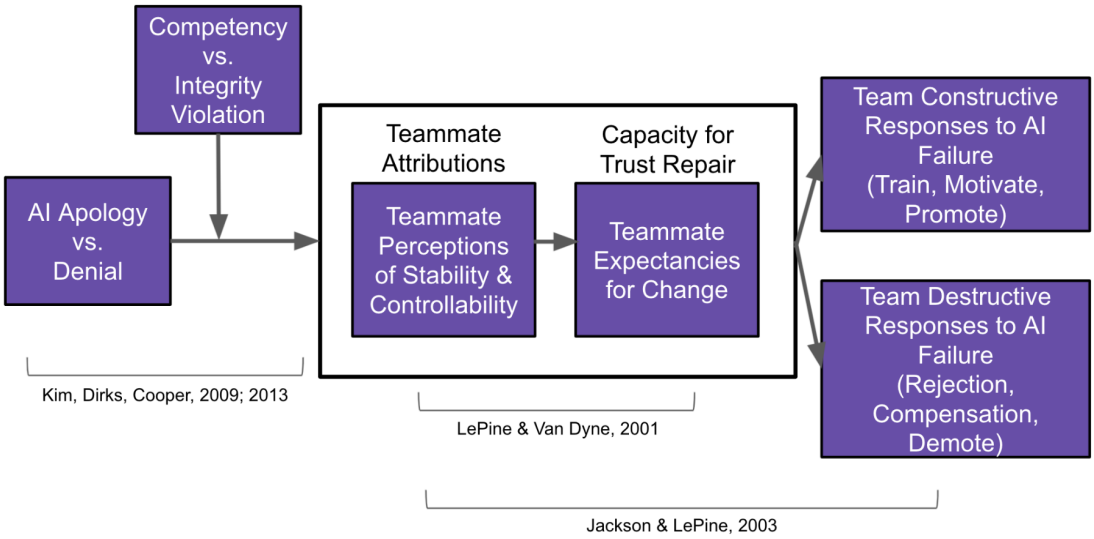
Cite	Behavior	Definition
De Visser et al., 2018	Ignore	Trustee deliberately ignores the occurrence of the trust violation.
De Visser et al., 2018	Deny	Trustee denies responsibility for the costly act.
-	Recognize	The trustee acknowledges that it performed a costly act.
Breazel, 2003; Riek et al., 2009; De Visser et al., 2009	Empathize	Trustee expresses empathy for the occurrence of the costly act.
De Visser et al., 2009; Jung, Martelaro, & Hinds, 2015	Emotionally Regulate	The trustee identifies the negative trigger (the event that violated the trustor's trust) and adds a normative statement to stay positive.
De Visser et al., 2009; Johnson et al., 2004; Kim et al., 2004, 2006	Blame	The trustee outwardly blames the trustor for the event that caused the violation of trust.
Dzindolet et al.,	Explain	The trustee provides an explanation for

2003		why they failed/violated the trustor's trust.
De Visser et al., 2009	Downplay	The trustee downplays the significance of the event that violated the trustor's trust.
Kim et al., 2004, 2006	Apology	A statement made by the trustee to the trustor to indicate they take responsibility.
-	Remorse	An emphasis in an apology that the trustee feels bad for breaking the trust of the trustor.
Robinette, Howard, & Wagner, 2015, 2017	Promise	A statement by the trustee that they will do better next time and/or make up for the current trust violation.
Jackson & LePine, 2003	Attribution: Controllability	An assessment from the trustor of the trustee as to whether or not the violation of trust was under the trustee's control.
Jackson & LePine, 2003	Attribution: Stability	An assessment from the trustor of the trustee as to whether or not the violation may be a one-off event or if the cause of the violation is something inherent to the trustee that is immutable.
Jackson & LePine, 2003	Attribution: Motivation	An assessment of the trustor of the trustee as to whether or not the trustee had malicious motivations or intent in violating their trust.
Jackson & LePine, 2003	Attribution: Ability	An assessment of the trustor of the trustee's capability of fulfilling their expectations.
Jackson & LePine, 2003	Attribution: Expectancy for change	An expectation of the trustor in the trustee as to whether or not the trustor can expect another violation in a similar situation or if the trustee is capable of changing their behavior in future circumstances.
Allemand, Amberg, Zimprich, & Fincham, 2007	Trait Forgiveness	A trait of the trustor that influences their likelihood to forgive.

-	Number of Prior Trust Violations	A weighted report built on the history of prior interactions.

As noted in a prior section of this dissertation, trust does not happen in a void… nor is it repaired in a void. Dyadic trust in the context of a team has been shown to have a galvanizing effect on trust reparation making it harder to regain the trust in a group (Kim, Cooper, Dirks, & Ferrin, 2013). This effect is mitigated by a correct matching of response type to violation type; for example, Kim and colleagues (2013) found that when a competence based trust violation was responded to with an apology by the trustee, rather than a denial, then trust was more likely to be repaired at both the dyadic level and group levels. And when trust was violated based on the trustee's integrity, a denial response from the trustee has a better chance of repairing trust than an apology. Kim and colleagues (2013) also found an order effect on when the assessment of trust happened, either the team or individual assessment happened first. They found that when the individual assessment happened first there was no effect on trust repair; however, when trust repair happened in the group first, individual assessments were affected. This finding demonstrates the galvanizing effect groups have on the trust repair process. Figure 4 depicts the proposed theoretical model for this dissertation.

Figure 4
Theoretical Model



Hypotheses and Justifications

In this dissertation I focus on the dyadic level trust repair between a human and an AI to build out a theoretical model of human-AI trust repair, and extend the theoretical model of peer attributions of poor performance to AI.

Merging two streams of academic research, the peer attributions of poor performers and trust repair, hypothesis 1 seeks to tease out and establish the ties between the type of trust violation, the attribution process, and the trust repair process. In human teams, the type of trust violation informs the effectiveness of the type of trustee response; this should hold for human-AI teams as well. However, the mechanism in which this works has remained unclear, attribution theory should fill in this gap. For either a competence based violation or an integrity based violation of trust, the observer will attribute greater stability to the AI's behavior when the AI responds with a denial than an apology because there is no signal for change.

Hypothesis 1: Following an AI trust violation, team members will attribute greater stability to the AI's behavior in response to a denial than in response to an apology.

Compared to a denial where the trustor will assess the stability of the trustee behavior to be stable and thus not likely to change, an apology will have the opposite effect. An apology following a trust violation will result in greater attributions of controllability as the trustee has signaled that they are aware of their actions and take responsibility.

Hypothesis 2: Following an AI trust violation, team members will attribute greater controllability to the AI's behavior in response to an apology than in response to a denial.

In traditional teams, the type of apology can affect the outcome depending on the type of trust violation. Following Kim and colleagues (2013) findings of mitigating effects of response type, Hypothesis 3 and 4 investigate this interaction by extending the model to human-AI teams as well as linking the type of trustee response to the peer attribution.

Hypothesis 3: Following an AI trust violation, team members will attribute lower stability to the AI's behavior in response to an apology if the violation was due to competence; however, if the violation was due to integrity, stability attributions will be higher with a denial as compared to an apology.

Hypothesis 4: Following an AI trust violation, team members will attribute greater controllability to the AI's behavior in response to an apology if the violation was due to competence; however, if the violation was due to integrity, controllability attributions will be higher with a denial as compared to an apology.

The attribution model includes behavioral outcomes of the peer attribution process. One of the two mechanisms that mediates the relationship between the attribution and the outcome behavior is the expectancy to change, or that the trustee will correct their behavior in future interactions. Following the attribution model for poor performance, the inverse relationship between controllability and stability, H5a and H5b posit that attributions of high controllability and low stability will positively predict high expectancy for change (H5A); and, inversely,

attributions of low controllability and high stability will negatively predict high expectancy to change.

Hypothesis 5a: Teammate attributions of controllability for AI failure are positively related to team members' expectancy for change.

Hypothesis 5b: Teammate attributions of stability for AI failure are negatively related to team members' expectancy for change.

Sympathy is the second mechanism that affects behavioral outcomes of the trustor. Sympathy as a mechanism in human AI trust repair may be the most interesting and unknown of the two mechanisms. It is entirely possible for sympathy to be a non-factor and drop out of the model entirely. But prior attribution theory research (Lepine & Van Dyne, 2001) indicates that there is an inverse relationship between sympathy and expectancy to change. That is, when the trustor is sympathetic toward the trustee who violated their trust, they are less likely to assume the behavior will change. Thus, H6 posits that this will function the same in human-AI trust repair as with interpersonal trust repair.

Hypothesis 6a: Teammate attributions of controllability for AI failure are negatively related to team members' sympathy.

Hypothesis 6b: Teammate attributions of stability for AI failure are positively related to team members' sympathy.

These hypotheses are tested in two studies. The first is an online vignette study conducted with a convenience sample on MTurk. The second study is a controlled experiment conducted at a university laboratory with a sample of students and community members. The measures used in both studies were the same. A series of factors related to the Covid-19 pandemic affected the ability to collect laboratory data, and so the first study was conducted in order to test relations in the model with adequate statistical power. The second study was conducted in the laboratory

with only 8 teams. Nonetheless, Study 2 enables a demonstration of the core relations in the context of an experienced team, and not a vignette team.

Study 1: Vignette Design

Procedure Study 1

In the first study I tested these hypotheses using a between subjects vignette study on a sample from MTurk. Participants will receive their pre-pre-measures survey before entering the study where I obtained consent. The pre measures are to obtain measures of bias before the participant interacts with any materials. The participants will then read a generic description of the AI and a human teammate followed by a brief measure of trust. The generic description will establish trust, and the measurement will confirm that this manipulation worked.

Following the manipulation check, participants were given a description of a scenario in which the AI or a person committed a trust violation in one of two ways, a competence based violation or an integrity based violation, followed by a manipulation check survey. This phase of study 1 establishes that the violation has truly broken trust.

In the final phase of study 1, participants read one two response behavior scenarios possible for each violation type, apology or denial; therefore there are four combinations a participant may encounter: 1) competence x apology; 2) competence x denial; 3) integrity x apology; and, 4) integrity x denial. A final measure of trust will be taken followed by the exit survey.

For example, the two scenarios, human or AI, read:

> "You are a member of a group in an upper level course. This course is part two-semester sequence. Your group is responsible for completing a major project that is worth half of your grade. Your team includes two other students and yourself. Successful

completion of the project requires equal contributions from all members, and therefore, each member of the group receives the same project grade."

"Your class is part of a pilot project to test a new educational technology named Vero. The university is poised to implement Vero in all team-based learning classes in the coming year. Vero is a fully-autonomous artificially intelligent robot representing state of the art technology. Vero can do [everything a human teammate can do with examples]. Your team includes one other student, Vero, and yourself. Successful completion of the project requires equal contributions from all members, and therefore, each member of the group receives the same project grade."

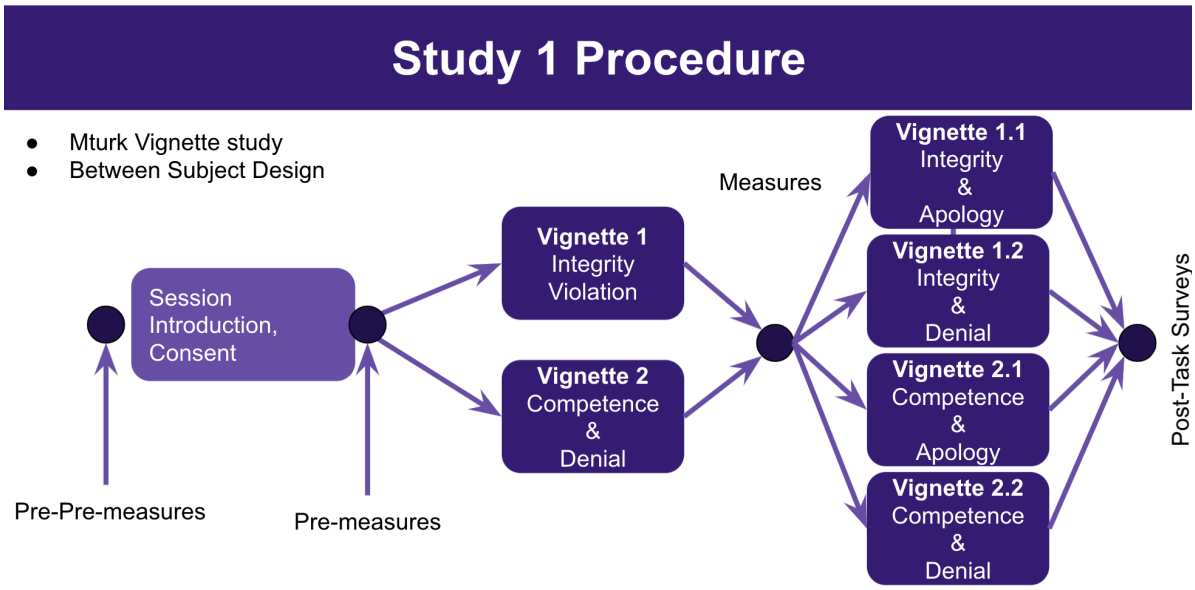
Participants were then given more information about the scenario in which the participants find out they have received a final letter grade of a B for the project. Upon further investigation the participant finds out that the AI or the human has submitted their own version of the project instead of what was agreed upon as a group. This trust violation is then clarified as to whether or not it was intentional (integrity) or accidental (competence) violations.

Participants were then given one of two forms of apology that are tailored to the violation type. For example the apology in the competence violation reads as follows.

"...You Slack Pat/Vero to ask about the project, and realized that the implementation of [the agreed upon idea] failed because Pat/Vero did not have adequate knowledge of the specific type of programming required to implement the feature set that was imagined in [the other group member's] design. You ask Pat/Vero about the implementation, and Vero/Pat accepted full responsibility, promised that they would not let it happen again, and reaffirmed a commitment to the group and its success. Vero/Pat said that you need not have any concerns about their competence next semester."

The full vignette study can be viewed in Appendix H. Figure 5 depicts the study procedure.

Figure 5
Procedure for Study 1



Study 1: Results

Below, I detail the results of my experimental studies of human-AI trust repair. First I present the results from Study 1, the vignette study, followed by Study 2 results from the laboratory study. Table 4 presents the descriptive statistics for Study 1 and Table 25 presents the same for Study 2.

Descriptives

Table 4 provides the descriptives for key variables at different time points. I report descriptives, including scale alphas, by the first factor.

Manipulation Check

While the study has two manipulation sources, the manipulation itself is the violation type (competence or integrity). In Table 5 and Figure, I report and show the results of the

manipulation check. This check was created by asking participants to rate the manipulation source (human or AI) on competence and integrity on a 5 point likert scale (very low on x, to very high on x). To test if participants were perceiving the manipulation correctly, I performed multiple one-way ANOVA tests, competence and integrity for each of the manipulation sources.

AI Integrity Manipulation

A 2x2 analysis of variance (ANOVA) was conducted to examine the effects of response type (apology, denial) and violation type (competence, integrity) on ratings of artificial intelligence (AI) integrity. The results showed a significant main effect for violation type ($F(1, 105) = 27.25$, $p < .001$), with higher ratings of AI integrity for competence violations ($M = 2.93$, $SE = 0.18$) compared to integrity violations ($M = 2.08$, $SE = 0.19$). The main effect for response type was not significant ($F(1, 105) = 3.64$, $p = .06$), nor was the interaction effect between response type and violation type ($F(1, 105) = 0.33$, $p = .57$).

A Bonferroni post-hoc test was conducted to explore pairwise comparisons of the two main effects. The results revealed significant differences in ratings of AI integrity between competence violations and integrity violations ($p < .001$), as well as between apology responses and integrity violations ($p = .002$), and between denial responses and integrity violations ($p < .001$).

AI Competence Manipulation

A 2x2 ANOVA was conducted to examine the effects of response type (apology, denial) and violation type (competence, integrity) on ratings of artificial intelligence (AI) competence. The results showed a significant main effect for violation type ($F(1, 105) = 19.18$, $p < .001$), with higher ratings of AI competence for integrity violations ($M = 3.46$, $SE = 0.21$) compared to competence violations ($M = 2.63$, $SE = 0.20$). The main effect for response type was not

significant ($F(1, 105) = 1.67$, $p = .20$), nor was the interaction effect between response type and violation type ($F(1, 105) = 0.19$, $p = .67$).

A Bonferroni post-hoc test was conducted to explore pairwise comparisons of the two main effects. The results revealed significant differences in ratings of AI competence between integrity violations and competence violations ($p < .001$), but no significant differences between apology and denial responses for either competence or integrity violations.

Furthermore, the contrast analysis showed a significant difference in AI competence ratings between competence violations and integrity violations for both apology responses ($p = .004$) and denial responses ($p < .001$).

Manipulation Check Summary

The results of the four ANOVAs above indicate that the manipulation was perceived as intended by participants. The results follow what one would expect when manipulating perceptions negatively, where competence is lower in the competence conditions and higher in the integrity conditions; and vice versa for ratings of integrity.

Hypothesis 1

An analysis of variance (ANOVA) was conducted to test the hypothesis that following an AI trust violation (Table 11), team members would attribute greater stability to the AI's behavior in response to a denial than in response to an apology. The results showed a significant main effect of response on stability attribution, $F(1, 103) = 5.80$, $p = .02$, with a partial eta squared effect size of .05.

Bonferroni post-hoc tests were conducted to determine the nature of the significant main effect, Table 12 & 13. The results of the post-hoc tests revealed that the mean stability attribution

following a denial (M = 3.69, SE = 0.10) was significantly lower than the mean stability attribution following an apology (M = 4.02, SE = 0.10), t(103) = -2.409, p = .02.

These results support Hypothesis 1, indicating that team members attribute greater stability to the AI's behavior in response to a denial than in response to an apology following a trust violation. Specifically, team members are more likely to view the AI's behavior as stable and predictable when the AI denies any wrongdoing than when it apologizes for the violation. The effect size was small, with the response explaining approximately 5% of the variance in stability attribution.

Hypothesis 2

Based on the results of the one-way ANOVA for Hypothesis 2, there was no significant difference in the attribution of controllability between an apology and a denial response following an AI trust violation, F(1, 103) = 2.28, p = .14. Therefore, Hypothesis 2 was not supported, see Table 14.

Hypothesis 3

A 2x2 ANOVA was conducted to examine the effects of violation type (competence vs. integrity) and response type (apology vs. denial) on stability attributions. The results revealed a significant main effect of violation type, F(1, 101) = 11.70, p < .001, ηp2 = .10, with lower stability attributions for competence violations compared to integrity violations. There was also a significant main effect of response type, F(1, 101) = 6.40, p = .01, ηp2 = .06, indicating that overall, team members attributed lower stability to the AI's behavior in response to an apology than a denial. Importantly, there was a significant interaction effect between violation type and response type, F(1, 101) = 6.16, p = .02, ηp2 = .06. Bonferroni post-hoc tests revealed that for competence violations, stability attributions were lower with an apology (M = 4.29, SE = 0.12)

than a denial (M = 3.85, SE = 0.14), t(101) = -3.10, p = .003, 95% CI [-0.95, -0.19]. In contrast, for integrity violations, stability attributions were higher with a denial (M = 3.52, SE = 0.14) than an apology (M = 3.72, SE = 0.13), although this contrast was not statistically significant, t(101) = -1.71, p = .09, 95% CI [-0.73, 0.06], see Table 15. These results offer partial support Hypothesis 3, indicating that the effect of response type on stability attributions depends on the type of violation committed.

While the interaction effect in the 2x2 ANOVA for Hypothesis 3 was not significant, it is still possible to explore whether the data support Hypothesis 3 based on the main effects of the ANOVA and the post-hoc tests.

The main effect of the violation (competence vs. integrity) was significant, suggesting that team members did attribute different levels of stability to the AI's behavior depending on whether the violation was due to competence or integrity. This result partially supports Hypothesis 3, which predicted lower stability attributions for an apology if the violation was due to competence. The post-hoc tests showed that team members attributed significantly lower stability to the AI's behavior in response to an apology when the violation was due to competence compared to integrity, which is consistent with Hypothesis 3.

The main effect of the response (apology vs. denial) was also significant, indicating that team members did attribute different levels of stability to the AI's behavior depending on whether the AI apologized or denied the violation. This result partially supports Hypothesis 3, which predicted that stability attributions would be higher with a denial than with an apology when the violation was due to integrity. The post-hoc tests did not show a significant difference between denial and apology conditions when the violation was due to integrity, so this part of Hypothesis 3 was not supported.

Therefore, while the interaction effect was not significant, the main effects and post-hoc tests offer partial support for Hypothesis 3.

Hypothesis 4

The results indicate a significant main effect of the Violation type on controllability attributions, $F(1, 101) = 71.68$, $p < .001$, with higher controllability attributions in the competence violation condition ($M = 2.47$, $SD = 0.64$) compared to the integrity violation condition ($M = 3.51$, $SD = 0.66$). However, the main effect of Response type on controllability attributions was only marginally significant, $F(1, 101) = 3.82$, $p = .05$, with higher controllability attributions for apologies ($M = 3.02$, $SD = 0.65$) compared to denials ($M = 2.52$, $SD = 0.68$), see Table 18.

The interaction effect was not significant, $F(1, 101) = 0.07$, $p = .8$, suggesting that the effect of the Response type on controllability attributions did not differ between the two Violation types.

Post-hoc analysis, Table 19 & 20, showed that for both competence and integrity violations, team members attributed higher controllability to the AI's behavior in response to a denial compared to an apology, all $ps < .001$.

Overall, these results provide partial support for Hypothesis 4, as the main effect of Response type on controllability attributions was only marginally significant and the interaction effect was not significant. However, the results do support the notion that team members attribute greater controllability to the AI's behavior following a denial compared to an apology, regardless of the type of violation.

Hypotheses 5a and 5b

A Pearson's correlation analysis was conducted to examine the relationship between teammate attributions of controllability for AI failure and team members' expectancy for change.

The correlation was not statistically significant (r = -0.08, p = .4). The 95% confidence interval for the correlation coefficient ranged from -0.27 to 0.12, indicating that there was no evidence of a significant positive relationship between these variables. Therefore, the results did not support Hypothesis 5a.

A Pearson's correlation was conducted to examine the relationship between teammate attributions of stability for AI failure and team members' expectancy for change. The correlation was significant and negative, r(103) = -0.69, p < .001, indicating that as team members' expectancy for change increased, teammate attributions of stability for AI failure decreased. The 95% confidence interval ranged from -0.78 to -0.57. These results provide support for Hypothesis 5b.

Hypotheses 6a and 6b

Based on the results of Pearson's correlation analysis, there was no significant relationship between teammate attributions of controllability for AI failure and team members' sympathy (r = 0.004, p = 0.97). The correlation coefficient was very small and not statistically significant, t(103) = 0.04, p = 0.97, indicating that there was no evidence to support hypothesis 6a. The 95% confidence interval ranged from -0.19 to 0.20, suggesting that the true correlation coefficient may be anywhere within this range with 95% confidence.

A Pearson’s correlation was conducted to examine the relationship between teammate attributions of stability for AI failure and team members’ sympathy. The correlation was not statistically significant, r(103) = 0.06, p = .55. The 95% confidence interval ranged from -0.14 to 0.25. Therefore, the results did not support Hypothesis 6b, which predicted a positive relationship between stability attributions and sympathy.

Study 1: Discussion

The present study aimed to examine how team members attribute stability and controllability to an AI's behavior in response to a trust violation, and how these attributions vary depending on the type of violation and response given by the AI. The findings provide insight into how people perceive and respond to AI behavior in situations where trust has been violated.

Hypothesis 1 predicted that team members would attribute greater stability to the AI's behavior in response to a denial than in response to an apology. The results showed a significant main effect of response on stability attribution, supporting Hypothesis 1. Team members attributed greater stability to the AI's behavior when it denied wrongdoing than when it apologized for the violation. The effect size was small, with the response explaining approximately 5% of the variance in stability attribution. These findings suggest that team members are more likely to view the AI's behavior as stable and predictable when the AI denies any wrongdoing than when it apologizes for the violation.

Hypothesis 2 predicted that team members would attribute greater controllability to the AI's behavior in response to an apology than in response to a denial. However, the results did not support Hypothesis 2. There was no significant difference in the attribution of controllability between an apology and a denial response following an AI trust violation.

Hypothesis 3 predicted that the effect of response type on stability attributions would depend on the type of violation committed. The results provided partial support for Hypothesis 3. The 2x2 ANOVA revealed a significant main effect of violation type, with lower stability attributions for competence violations compared to integrity violations. There was also a significant main effect of response type, with overall lower stability attributions to the AI's behavior in response to an apology than a denial. The interaction effect between violation type and response type was significant, indicating that the effect of response type on stability

attributions depends on the type of violation committed. Specifically, team members attributed significantly lower stability to the AI's behavior in response to an apology when the violation was due to competence compared to integrity. However, there was no significant difference between denial and apology conditions when the violation was due to integrity. Therefore, the main effects and post-hoc tests suggest partial support for Hypothesis 3.

Hypothesis 4 predicted that team members would attribute greater controllability to the AI's behavior in response to a competence violation compared to an integrity violation. The results showed a significant main effect of violation type on controllability attributions, with higher controllability attributions in the competence violation condition compared to the integrity violation condition. However, the main effect of response type on controllability attributions was only marginally significant.

Overall, the findings suggest that team members' attributions of stability and controllability to an AI's behavior in response to a trust violation depend on the type of response given by the AI and the type of violation committed. When an AI denies wrongdoing, team members are more likely to view its behavior as stable and predictable, whereas an apology may lead to lower stability attributions. The results also highlight the importance of considering the type of violation committed, with competence violations leading to lower stability and higher controllability attributions compared to integrity violations.

Study 1: Limitations

Limitations of the study include the use of hypothetical scenarios and a vignette-simulated AI, which may not fully capture the complexity of real-life trust violations involving AI. Furthermore, vignette studies may provide cognitively distant stimuli with lower fidelity. Future research could use real-world scenarios to examine how team members attribute

stability and controllability to AI behavior in response to trust violations. Additionally, the study only examined one aspect of team members' responses to AI trust violations, and future research could explore other factors such as trust repair strategies and their effectiveness in restoring trust in AI. This includes new manipulations that directly target factors like sympathy and controllability.

Study 2: Laboratory Experiment

Based on feedback and edits suggested by Study 1, I then tested the focal hypotheses in a laboratory study. Participants were recruited from the university subject pools, including both students and community members, and came into the ATLAS lab space. Once in the lab, participants were given consent, the pre-pre-measures, shown a video introducing the AI they worked with during the experiment, and then given the pre-measures which contain a manipulation check to ensure trust has been established and to measure baseline trust.

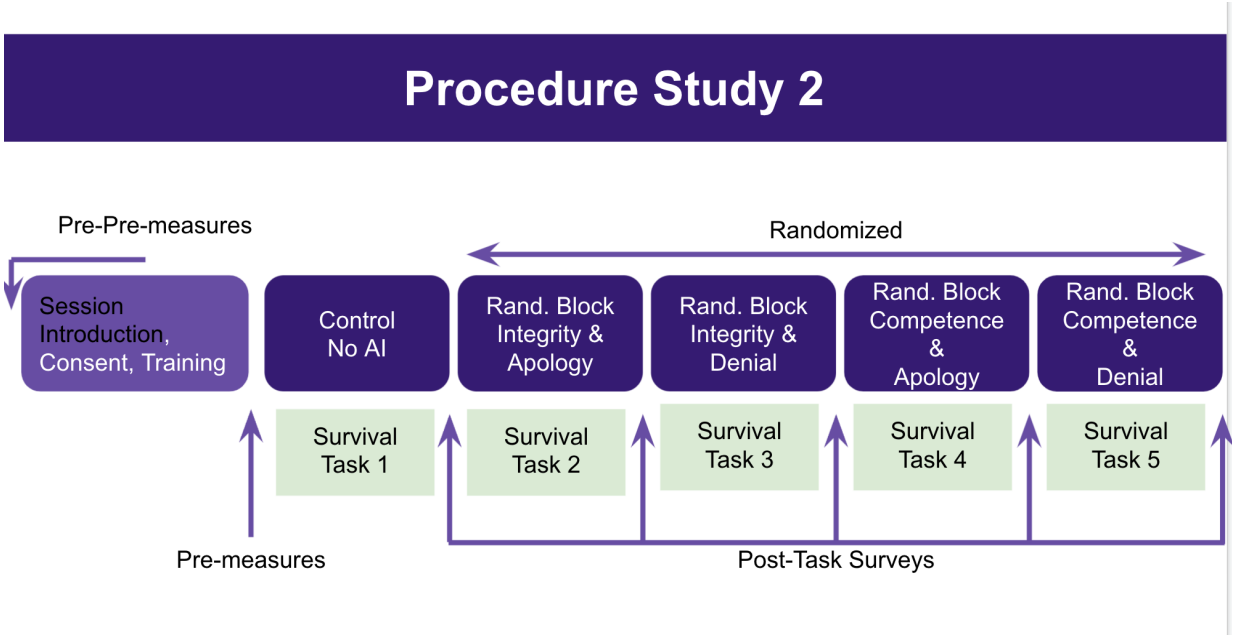
Study 2: Method

This experiment was part of a larger study investigating the neurological markers of synchrony and team process in Human-AI teams, and so for 4 of the teams, each member and the confederate AI were outfitted with Advanced Brain Monitoring (ABM) devices throughout the study. Four teams were not outfitted with ABMs.

Participants then began the first task, the control task. In this task participants will work together without an AI teammate to establish trust with the team. After the control task, participants will then team with the AI confederate to complete a randomized series of survival tasks. Each task, including the control task required participants to participate via our survival task platform. When participants first log into the platform they are prompted to take a profile picture. The platform then randomizes turns for each task and participants have an allotted time

to make the change they want to see to the rankings, then the turn will advance to the next participant. Once all participants, including the AI confederate, completed their turn the round ends and the chat opens for a period of time. During this time the participants can discuss what happened in the prior round and make plans for future rounds. There are four rounds per task and five tasks total in the study. Once the rounds for the task finished, participants were given a short set of surveys to assess attributions and trust, and then they moved on to the next survival task. Figure 5 depicts the procedure for study 2.

Figure 6
Procedure for Study 2



Measures

OTI

Eight items from the Organizational Trust Inventory (OTI-S) (Cummings & Bromiley, 1996), adapted to the individual level, rather than the organizational level, will be used to measure trust dyadically. Example items include, "My level of confidence that ____ is

technically competent at the critical elements of his or her job is ____," and "My level of confidence that ____ is able to do his or her job in an acceptable manner is ____." Participants will rate their trust in their teammates using a seven-point Likert scale, 1 - Nearly zero, 2 - Very low, 3 - Low, 4 - 50-50, 5 - High, 6 - Very high, 7 - Near 100%.

Attribution scales

Peer attributions will be assessed using assorted scales from Jackson & Lepine (2003). These scales consist of seven subscales, one for each factor in the peer attribution model, Ability (e.g. "___ has the ability to perform well."), motivation (e.g. ____ tries hard to complete group tasks."), compliance (e.g. "____ is open to the opinions of others."), controllability (e.g. "The cause of ____'s low performance was something that is controllable by _____."), stability (e.g. "____'s performance is temporary."), expectancy for change (e.g. "_____ can become a better performer."), and sympathy (e.g. "I feel sympathy towards ____."). Participants will rate themselves on a five-point Likert scale of 1 (Strongly Disagree) to 5 (Strongly Agree).

Study 2: Lab study

In order to address the limitation of a vignette study, I followed up the findings with a laboratory study. As described in detail in the method section, in Study 2, individuals were responding to experienced violations and responses in a short-term laboratory team while interacting with a wizard of oz AI. A total of 8 teams and 21 individuals were observed in Study 2.

Study 2: Results

Descriptives for Study 2

A total of 21 participants (N = 21) completed the study, and various measures were assessed for internal consistency using Cronbach's alpha. The AI Organizational Trust Inventory - Team scale consisted of 8 items and had a Cronbach's alpha of 0.94, indicating high internal

consistency. The mean score for this scale was 5.4 (SD = 0.87). The AI Organizational Trust Inventory - Dyadic scale included 8 items, with a Cronbach's alpha of 0.9, suggesting good internal consistency, and a mean score of 4.2 (SD = 1.2).

The AI Ability measure, consisting of 3 items, demonstrated a Cronbach's alpha of 0.96, indicating high internal consistency, and a mean score of 3.8 (SD = 1). The AI Motivation scale included 3 items and had a Cronbach's alpha of 0.89, suggesting good internal consistency, with a mean score of 3.6 (SD = 1.1). The AI Compliance measure consisted of 3 items, with a Cronbach's alpha of 0.92, indicating high internal consistency, and a mean score of 2.4 (SD = 1.1). The AI Stability scale, including 4 items, had a Cronbach's alpha of 0.88, suggesting good internal consistency, and a mean score of 3.5 (SD = 0.91). The AI Controllability measure, consisting of 6 items, demonstrated a Cronbach's alpha of 0.92, indicating high internal consistency, and a mean score of 3.1 (SD = 1.1). The AI Sympathy scale included 3 items, with a Cronbach's alpha of 0.89, suggesting good internal consistency, and a mean score of 2.0 (SD = 0.96). Lastly, the AI Expectancy to Change measure, consisting of 3 items, had a Cronbach's alpha of 0.94, indicating high internal consistency, and a mean score of 4.3 (SD = 0.63).

The main difference between Study 1 and Study 2 is that Study 2 utilizes a mixed experimental design where the response behavior factor is within subjects and the violation type factor is between subjects. This required a few changes to the inferential statistics used for analysis. Mainly, when looking at the effects of manipulations, the mixed design needed to be accounted for.

A mixed two-way ANOVA was conducted to examine the effects of response behavior (denial vs. apology) and violation type (competence vs. integrity) on stability attributions. The

results indicated no significant main effect of response behavior, $F(1, 35) = .13$, $p = .72$, $\eta^2 = .004$, or violation type, $F(1, 35) = 2.91$, $p = .10$, $\eta^2 = .08$, nor a significant interaction between the two factors, $F(1, 35) = .67$, $p = .41$, $\eta^2 = .02$. Therefore, Hypothesis 1 and Hypothesis 3 were not supported.

Another mixed two-way ANOVA was conducted to examine the effects of response behavior (denial vs. apology) and violation type (competence vs. integrity) on controllability attributions. The results indicated no significant main effect of response behavior, $F(1, 35) = .15$, $p = .70$, $\eta^2 = .004$, or violation type, $F(1, 35) = 1.92$, $p = .18$, $\eta^2 = .05$, nor a significant interaction between the two factors, $F(1, 35) = .00$, $p = .96$, $\eta^2 < .001$. Therefore, Hypothesis 2 and Hypothesis 4 were not supported.

Pearson's correlation analysis revealed a significant negative relationship between stability attributions and expectancy for change, $r(39) = -.46$, $p = .003$, 95% CI [-.67, -.17]. This result supports Hypothesis 5b. However, there was no significant relationship between controllability attributions and expectancy for change, $r(39) = .15$, $p = .37$, 95% CI [-.17, .43], failing to support Hypothesis 5a.

In addition, Pearson's correlation analysis showed no significant relationship between controllability attributions and sympathy, $r(39) = -.09$, $p = .59$, 95% CI [-.39, .23], failing to support Hypothesis 6a. Similarly, there was no significant relationship between stability attributions and sympathy, $r(39) = -.09$, $p = .59$, 95% CI [-.39, .23], failing to support Hypothesis 6b.

Study 2: Discussion

The present study investigated the effects of response behavior (denial vs. apology) and violation type (competence vs. integrity) on stability and controllability attributions following an

AI trust violation. Additionally, the study examined the relationships between these attributions and team members' expectancy for change and sympathy. The findings of the study did not support the hypotheses regarding the main effects and interactions of response behavior and violation type on stability and controllability attributions. Furthermore, only one of the four hypotheses concerning the relationships between attributions and expectancy for change and sympathy was supported.

Contrary to Hypothesis 1, the results indicated that there was no significant difference in stability attributions between denial and apology conditions. This suggests that team members did not perceive the AI's behavior as more stable following a denial than an apology, which challenges the notion that denial might be associated with greater perceived stability. Similarly, Hypothesis 2 was not supported, indicating that there was no significant difference in controllability attributions between apology and denial conditions. This finding contradicts the expectation that team members would attribute greater controllability to the AI's behavior in response to an apology compared to a denial.

The lack of significant interactions between response behavior and violation type on stability and controllability attributions resulted in the non-support of Hypothesis 3 and Hypothesis 4. These results suggest that the influence of response behavior on attributions does not differ depending on the type of trust violation. This finding is surprising, as it was expected that different trust violation types (competence vs. integrity) would result in different patterns of attributions in response to denial or apology.

In terms of the relationships between attributions and expectancy for change and sympathy, only Hypothesis 5b was supported. The significant negative relationship between stability attributions and expectancy for change indicates that as team members attributed greater

stability to the AI's behavior, they had lower expectations for change. This finding is in line with previous research, which has demonstrated that stability attributions can negatively influence the expectation of change in future behavior (e.g., Weiner, 1985). However, Hypothesis 5a, which proposed a positive relationship between controllability attributions and expectancy for change, was not supported. This result implies that team members' attributions of controllability for AI failure may not significantly influence their expectancy for change.

Lastly, neither Hypothesis 6a nor Hypothesis 6b was supported, as there were no significant relationships between controllability or stability attributions and sympathy. These findings suggest that team members' sympathy towards the AI might not be directly related to their attributions of stability and controllability following trust violations. This result contradicts previous research, which has shown that attributions can impact sympathy and other affective responses (Weiner, 1985).

The lack of support for the hypotheses in this study could be attributed to several factors. One possibility is that participants might have responded differently to AI trust violations compared to human trust violations, as previous research has mainly focused on human-human trust dynamics. Another explanation could be related to the specific context of the study or the nature of the AI system and its violations, which might not have elicited the expected attributions. Finally, the relatively small sample size might have limited the study's statistical power to detect significant effects.

Despite the non-support of most hypotheses, this study has important theoretical and practical implications. First, it highlights the need to further investigate the factors that influence attributions in the context of AI trust violations. Future research could explore other variables that may impact attributions, such as the severity of the violation, the history of the AI's

performance, or the individual characteristics of team members. Moreover, it could be valuable to examine the role of different types of apologies and denials in shaping attributions, as well as the potential moderating effects of trust repair strategies. Second, this study underscores the importance of understanding the relationships between attributions, expectancy for change, and sympathy in the context of AI trust violations, as these factors can influence team members' perceptions, emotions, and willingness to work with AI systems.

From a practical standpoint, the findings of this study can inform the development of AI systems and their interactions with human team members. By understanding the factors that shape attributions and their consequences, AI designers and engineers can develop more effective trust repair strategies in the event of trust violations. Additionally, insights from this study can help organizations better manage human-AI team dynamics, facilitating more productive collaboration and improving overall team performance.

There are several limitations to this study that should be acknowledged. First, as mentioned earlier, the relatively small sample size may have limited the study's statistical power to detect significant effects. Future research with larger sample sizes could provide more robust findings. Second, the study's experimental design may not have fully captured the complexity of real-world human-AI interactions, which can involve multiple trust violations, varying levels of interdependence, and diverse tasks. Further research using more ecologically valid settings and designs could help to address this limitation.

In conclusion, the present study sheds light on the complex interplay between response behavior, violation type, and attributions in the context of AI trust violations. Although most hypotheses were not supported, the findings provide valuable insights into the factors that may influence attributions and their relationships with expectancy for change and sympathy. This

study underscores the importance of further research on this topic, as understanding the nuances of human-AI trust dynamics can contribute to the development of more effective AI systems and improved collaboration between humans and AI in various domains.

Study 2: Limitations

Several limitations of the present study should be acknowledged. First, the relatively small sample size (N = 58) may have limited the study's statistical power to detect significant effects. Future research with larger sample sizes could provide more robust findings. Second, the study's experimental design may not have fully captured the complexity of real-world human-AI interactions, which can involve multiple trust violations, varying levels of interdependence, and diverse tasks. Further research using more ecologically valid settings and designs could help to address this limitation.

Additionally, the lack of support for the hypotheses in this study could be attributed to factors such as participants potentially responding differently to AI trust violations compared to human trust violations, as previous research has mainly focused on human-human trust dynamics. Another explanation could be related to the specific context of the study or the nature of the AI system and its violations, which might not have elicited the expected attributions.

These limitations highlight the need for future research to further investigate the factors that influence attributions in the context of AI trust violations. For instance, exploring other variables that may impact attributions, such as the severity of the violation, the history of the AI's performance, or the individual characteristics of team members, could provide a more comprehensive understanding of the nuances of human-AI trust dynamics.

Conclusion: Study 1 and Study 2

The two studies presented in this paper aimed to investigate the effects of response behavior (denial vs. apology) and violation type (competence vs. integrity) on stability and controllability attributions following AI trust violations. Additionally, the relationships between these attributions and team members' expectancy for change and sympathy were examined. While some hypotheses were supported in Study 1, most hypotheses in Study 2 were not supported. Despite these mixed findings, the studies provide valuable insights into the factors that may influence attributions and their relationships with expectancy for change and sympathy in the context of AI trust violations.

The results of both studies underscore the importance of understanding the factors that shape attributions and their consequences in the context of AI trust violations. By understanding the nuances of human-AI trust dynamics, AI designers, engineers, and organizations can develop more effective trust repair strategies and better manage human-AI team dynamics. This could ultimately lead to more productive collaboration and improved team performance in various domains.

An important area for future research is the investigation of the moderating effects of stability and controllability attributions on expectancy for change. While the current studies provided some insights into the relationships between attributions and expectancy for change, a more nuanced understanding of these relationships is needed. In particular, it is possible that the effects of stability and controllability attributions on expectancy for change might not be linear or direct, but rather moderated by various factors, such as the nature of the trust violation, the AI's past performance, or individual differences among team members.

Examining the moderating effects of stability and controllability attributions on expectancy for change could help clarify the conditions under which these attributions are more or less influential in shaping team members' expectations regarding AI behavior. For instance, it could be that stability attributions have a stronger impact on expectancy for change when trust violations are severe or when the AI has a history of consistent performance. Similarly, controllability attributions might be more influential in shaping expectancy for change when team members perceive the AI as having a high level of autonomy or when they have a high degree of interdependence with the AI system.

In addition to the factors already discussed, it is essential to consider the role of cultural differences in shaping attributions and responses to AI trust violations. Culture can influence individuals' interpretations of events, their expectations regarding the behavior of AI systems, and their reactions to trust violations. For example, individuals from collectivist cultures may place a higher emphasis on the role of the group in shaping AI behavior and may be more inclined to attribute trust violations to external factors, whereas those from individualistic cultures may be more likely to focus on the AI's internal characteristics. Understanding these cultural differences and their implications for attributions and trust repair strategies is vital for the development of AI systems that can effectively function in diverse cultural contexts.

Furthermore, it is important to consider the ethical implications of AI trust violations and the development of trust repair strategies. As AI systems continue to become more integrated into various aspects of human life, the potential consequences of trust violations become more significant, and the ethical considerations surrounding the design and implementation of these systems become increasingly complex. For instance, AI designers and engineers must balance the need for transparency in AI systems with the desire to protect users' privacy and maintain

system security. Additionally, the development of trust repair strategies may involve ethical dilemmas, such as whether an AI system should prioritize repairing trust with individual users or with society as a whole.

Another area of inquiry that warrants further exploration is the potential impact of AI trust violations on users' well-being and mental health. The increasing prevalence of AI systems in various aspects of human life, including healthcare, education, and social interactions, may lead to the development of strong emotional bonds between humans and AI. Consequently, AI trust violations could have a significant impact on users' emotions, psychological well-being, and mental health. Understanding the potential psychological consequences of AI trust violations and developing strategies to mitigate these effects is crucial for ensuring that AI systems not only function effectively but also promote human well-being.

Lastly, it is critical to recognize the potential impact of AI trust violations on society as a whole. As AI systems become more integrated into various domains, trust violations could lead to significant consequences, such as reduced public trust in AI systems, regulatory backlash, or even widespread social unrest. Therefore, it is essential for researchers, AI developers, and policymakers to work collaboratively to establish guidelines and best practices for AI trust management, ensuring that AI systems are developed and deployed responsibly and that potential trust violations are addressed in a timely and effective manner.

Several limitations in the current research should be acknowledged, including the relatively small sample sizes and the experimental designs, which may not have fully captured the complexity of real-world human-AI interactions. Future research could address these limitations by employing larger sample sizes and more ecologically valid settings and designs.

Moreover, future research should continue to explore the factors that influence attributions in the context of AI trust violations, such as the severity of the violation, the history of the AI's performance, and individual characteristics of team members. Examining the role of different types of apologies and denials, as well as potential moderating effects of trust repair strategies, could further enrich our understanding of human-AI trust dynamics.

In conclusion, although the findings from the two studies were mixed, they contribute to our understanding of the complex interplay between response behavior, violation type, and attributions in the context of AI trust violations. These insights, along with the exploration of the moderating effects of stability and controllability attributions on expectancy for change, and the consideration of cultural differences, ethical implications, users' well-being, and societal impact, can inform the development of more effective AI systems and promote improved collaboration between humans and AI across various domains.

Appendix A: Theoretical Tables

Table 1

Trusting Relationships Possible in a Mixed Initiative Team

Interaction Type	Level	Trustor(s)	Trustee(s)
Interpersonal	Dyadic	Human	A person that will be or is the cause of an event.
Human-AI Trust	Dyadic	Human	An autonomous artificial intelligence in any form of embodiment that will be the cause of an event.
Team Trust	Team	Human	Other members of the team as trustees lead to the emergence of the team as a whole becoming the trustee.
Trust in Human-Autonomy Teaming	Team	Human	Other members of the team, including a dyadic relationship between the trustor and an AI teammate trustee, lead to the emergence of the team as a whole becoming the trustee.

Table 2
Exemplar Definitions of Trust

Citation	Trust Type	Definition
Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709-734.	Interpersonal Trust	"[Trust is] the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other part (p. 217)."
McAllister, D. J. (1995). Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal*, 38(1), 24-59.	Interpersonal Trust	"Interpersonal trust is the extent to which a person is confident in, and willing to act on the basis of, the words, actions, and decisions of another. (p. 25)"
Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50-80.	Human-Machine Trust	"A simple definition of trust [...] is the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability. (p. 54)"
Fulmer, C. A., & Gelfand, M. J. (2012). At what level (and in whom) we trust: Trust across multiple organizational levels. *Journal of Management*, *38*(4), 1167-1230.	Team Trust	"[Team trust is] a shared psychological state among team members comprising willingness to accept vulnerability based on positive expectations of a specific other or others. (p. 1174)"

Table 3

New and Related Constructs

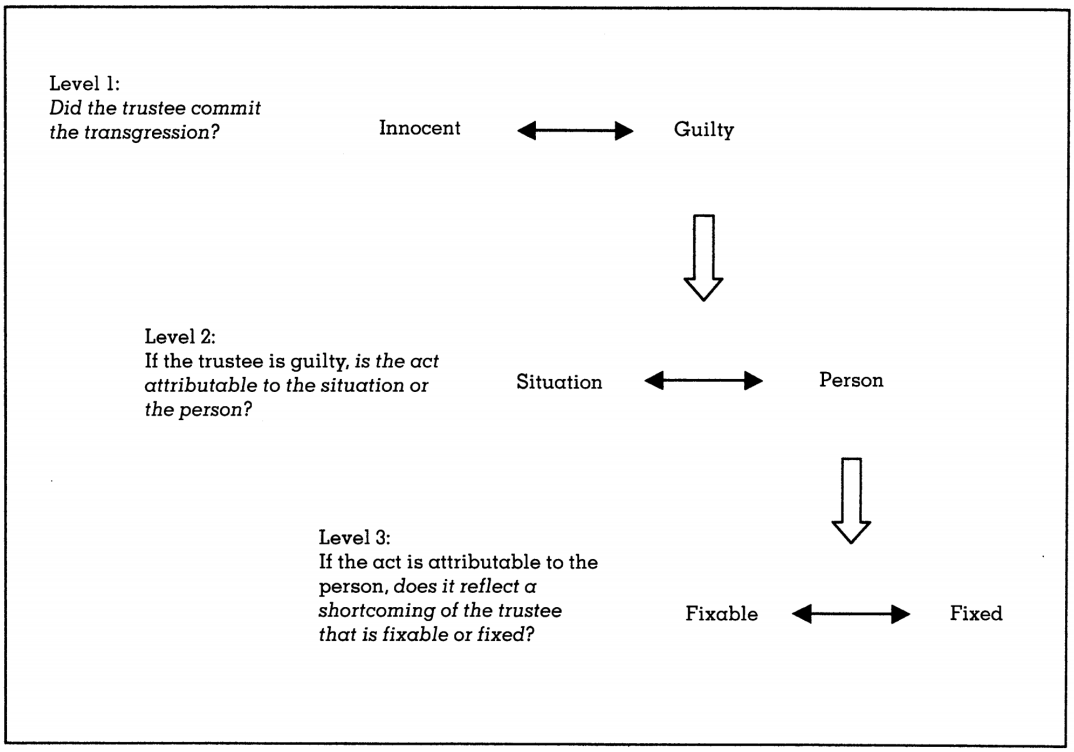
Author(s)	Construct (State/Trait)	Definition
Mayer, Davis, & Schoorman, 1995	Trustworthiness	Attributes of the trustee perceived by the trustor. Composed of: ability, integrity, and benevolence.
Gills, Bois, Finegan, & McNally	Propensity to Trust	The trustor's unique likelihood of trusting a trustee.
Hanoch, Johnson, & Wilke, 2006	Propensity for Risk Taking Behavior (PRTB)	A combination of traits in the trustor that moderate the likelihood of accepting risk.
Tomlinson & Mayer, 2009	Trust Repair	The period of time and process after the trustor's trust has been violated where the trustor must decide to trust or not to trust the trustee again.
Mayer, Davis, & Schoorman, 1995	Trust	"[Trust is] the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other part (p. 217)."
--	Trust Establishment	The front end of the trust cycle. Trust establishment is moderated by propensity to trust and the perceived trustworthiness of the trustee.
Enright, Freedman, & Rique, 1998	Forgiveness	An unconditional, volitional response to another person.
Elangovan, Auer-Rizzi & Szabo, 2006	Trust Erosion	The factors that influence the amount of trust lost after a trust violation, grounded in

		the trustor's knowledge of the trustee's prior performance. Trust erosion assumes a state of high initial trust.
Elangovan & Shapiro, 1998	Trust Violation	Trust is violated when the trustor perceives the trustee as acting in a way that does not fulfill their expectations.
Hirschman, 1984	Trust Depletion	The slow decline of trust due to little or no contact between the trustee and the trustor for an extended period of time.
Fincham, 2000	Reconciliation	Reconciliation is a bilateral process that requires goodwill by both the trustee and the trustor after a trust violation has occurred.

Appendix B: Theoretical Figures

Figure 1

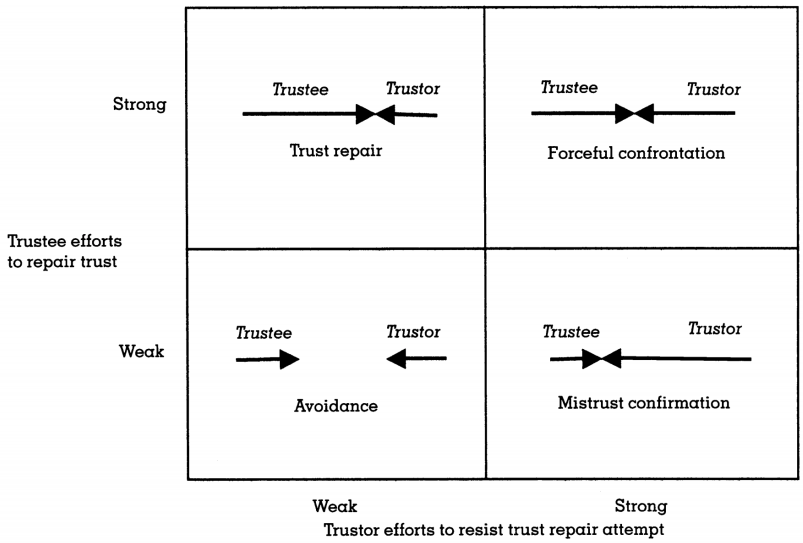
Bilateral Model of Trust Repair



Note: Model from Kim, Dirks, and Cooper's (2009) paper on trust repair.

Figure 2

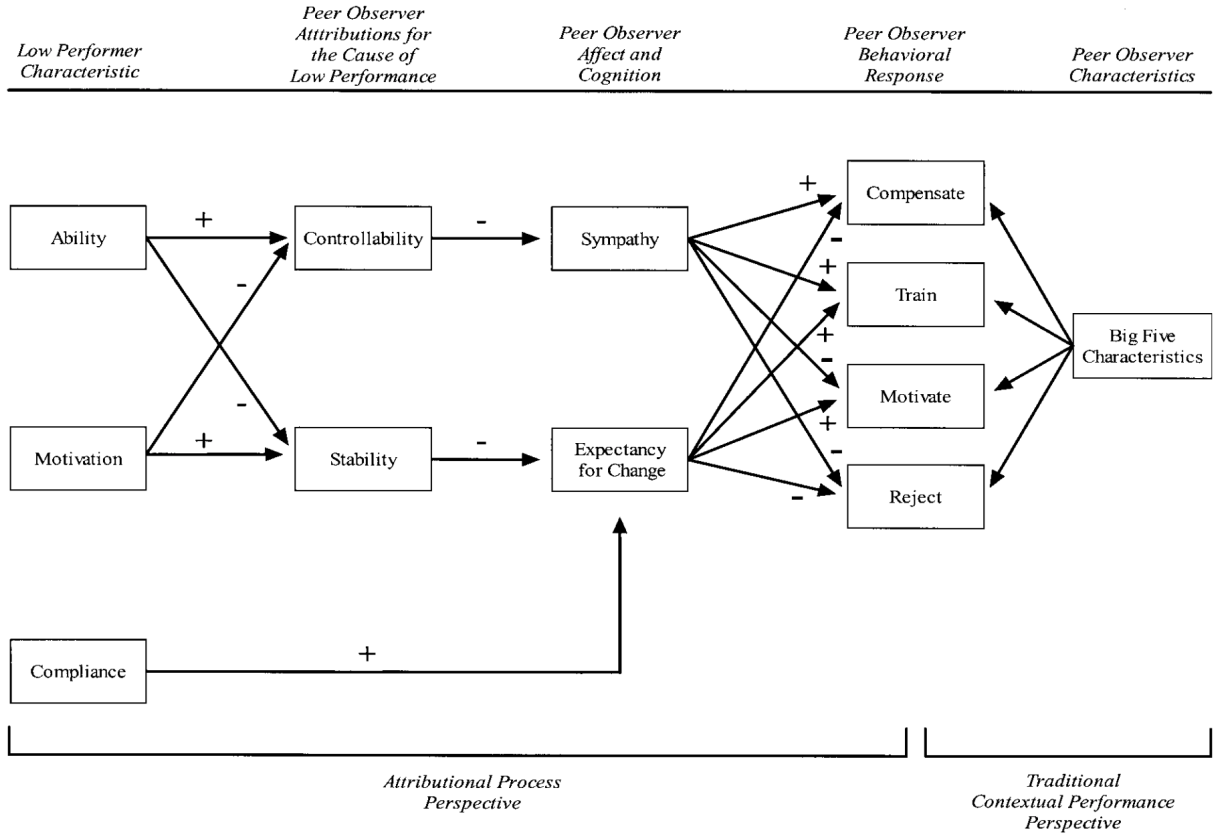
Potential Outcomes of Negotiation Efforts by a Trustor and Trustee



Note: Contingency table from Kim, Dirks, and Cooper's (2009) paper on trust repair.

Figure 3

Attributional Model of Peer Poor Performance



Note: Model from Jackson and LePine's (2003) paper on perceptions of peer poor performance.

Figure 4

Proposed Theoretical Model

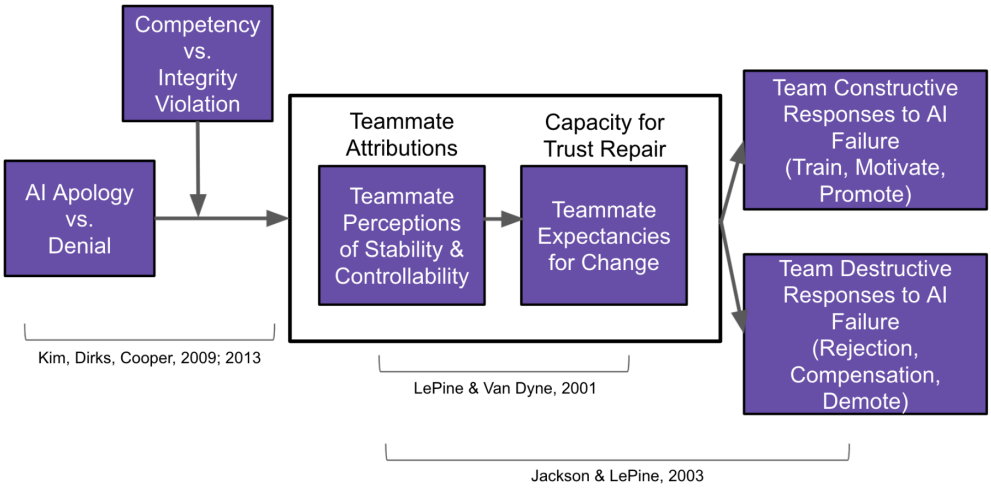


Figure 5

Procedure for Study 1

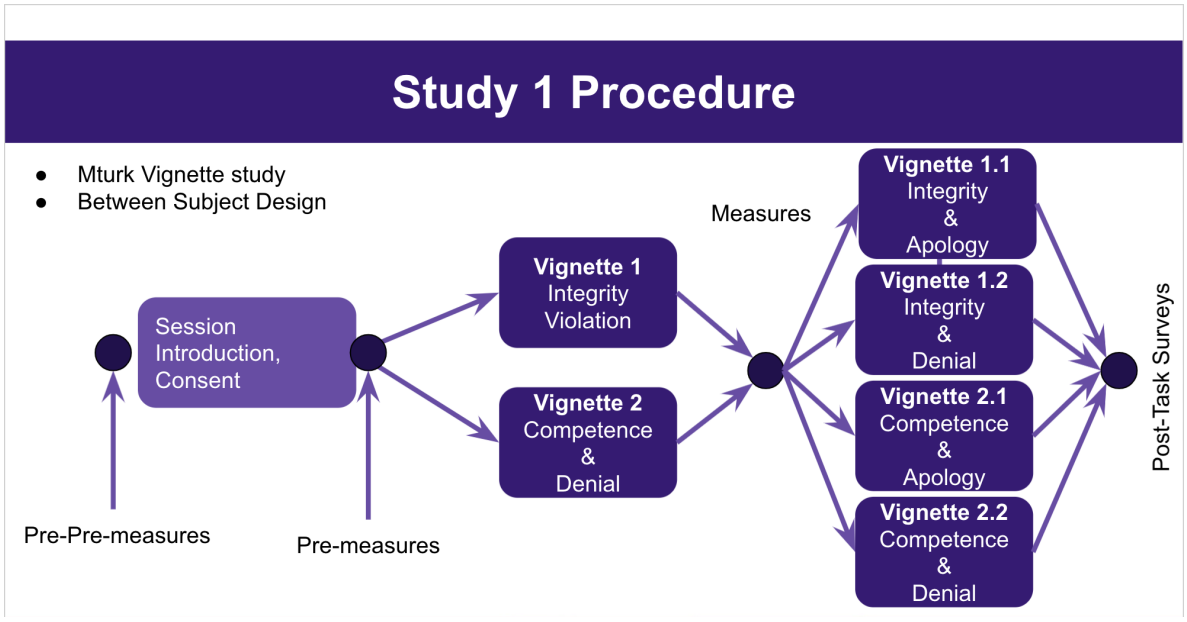
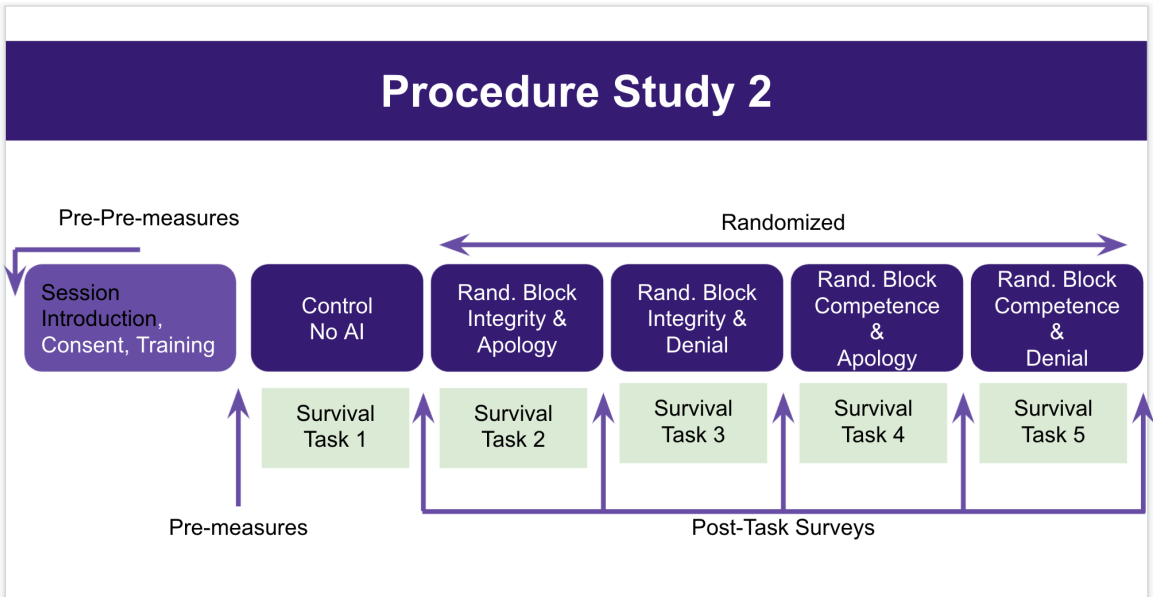


Figure 6

Procedure for Study 2



Appendix C: Organizational Trust Inventory (OTI) - Dyadic (AI) Survey

Cite: Cummings, L. L., & Bromiley, P. (1996). The organizational trust inventory (OTI). *Trust in organizations: Frontiers of theory and research*, *302*(330), 39-52.
Nyhan, R. C., & Marlowe Jr, H. A. (1997). Development and psychometric properties of the organizational trust inventory. Evaluation Review, 21(5), 614-635.

Instructions: Complete each of the following statements on a scale of 1 - Nearly Zero to 7 - Near 100%.. After reading the statement, select the number from the scale provided that is closest to your opinion that appropriately fills in the second blank at the end of the statement.

1 - Nearly zero, 2 - Very low, 3 - Low, 4 - 50-50, 5 - High, 6 - Very high, 7 - Near 100%

Assuming you have access to the AI in the future….

1. My level of confidence that the AI is technically competent at the critical elements of his or her job is ____
2. My level of confidence that the AI makes well thought out decisions about his or her job is ____
3. My level of confidence that the AI follows through on assignments is ___
4. My level of confidence that the AI has an acceptable level of understanding of his/her job is ___
5. My level of confidence that the AI is able to do his or her job in an acceptable manner is ____
6. When the AI tells me something, my level of confidence that I can rely on what he/she tells me is ____
7. My confidence in the AI to do the job without causing other problems is ____
8. My level of confidence that the AI thinks through what he or she is doing on the job is _____

Appendix D: Assorted Attribution Scales

Jackson, C. L., & LePine, J. A. (2003). Peer responses to a team's weakest link: A test and extension of LePine and Van Dyne's model. *Journal of Applied Psychology*, *88*(3), 459.

Assessed using a 5-point Likert scale with anchors of 1 (strongly disagree) and 5 (strongly agree).

Ability

1. ____ has the ability to perform well.
2. ___ has the capacity to be highly effective.
3. ____ is capable of being a good performer.

Motivation

1. ____ is motivated to do well.
2. ____ tries hard to complete group tasks.
3. ____ works hard to achieve group goals.

Compliance

1. ____ is open to the opinions of others.
2. ____ is receptive to others' ideas.
3. ____ listens to the suggestions of others.

Controllability

1. The cause of ____'s low performance was something that is controllable by _____.
2. The cause of ____'s low performance was something that was intended by ____.
3. ____ is not responsible for the cause of ____'s low performance. (R)
4. ____ is responsible for the cause of ____'s low performance.
5. ____'s low performance is not intended by ___. (R)
6. The cause of ____'s low performance is not controllable by ____. (R)

Stability

1. ____’s performance is temporary.
2. ____’s performance is permanent. (R)
3. ____’s performance is changeable.
4. ____’s performance is unchangeable. (R)

Expectancy for Change

1. _____ can become more effective.
2. _____ can become a better performer.
3. _____ performance can improve.

Sympathy

1. I feel sympathy towards ____.
2. I feel pity towards ____.
3. I feel concern for ____.

Appendix E: Organizational Trust Inventory (OTI-S) - Team Survey

Cite: Cummings, L. L., & Bromiley, P. (1996). The organizational trust inventory (OTI). *Trust in organizations: Frontiers of theory and research*, *302*(330), 39-52.
Nyhan, R. C., & Marlowe Jr, H. A. (1997). Development and psychometric properties of the organizational trust inventory. *Evaluation Review*, *21*(5), 614-635.

Instructions**:** Complete each of the following statements on a scale of 1 - Nearly Zero to 7 - Near 100%.. After reading the statement, select the number from the scale provided that is closest to your opinion that appropriately fills in the second blank at the end of the statement.

1 - Nearly zero, 2 - Very low, 3 - Low, 4 - 50-50, 5 - High, 6 - Very high, 7 - Near 100%

1. My level of confidence that this team is technically competent at the critical elements of their respective jobs is _____.
2. My level of confidence that my team will make well thought out decisions about their respective jobs is _____.
3. My level of confidence that my team will follow through on assignments is ____.
4. My level of confidence that my team has an acceptable level of understanding of their respective jobs is ____.
5. My level of confidence that my team will be able to do their respective jobs in an acceptable manner is ____.
6. When my team tells me something, my level of confidence that I can rely on what they tell me is ____.
7. My confidence in my team to do the job without causing other problems is ____.
8. My level of confidence that my team will think through what they are doing on the job is ____.

Appendix F: Vignettes for Study 1

Intro for Human Condition

You are a member of a group in an upper level course. This course is part two-semester sequence. Your group is responsible for completing a major project that is worth half of your grade. Your team includes two other students, Pat and Alex, and yourself. Successful completion of the project requires equal contributions from all members, and therefore, each member of the group receives the same project grade.

One day each week the group meets to work on the project. As the semester draws to a close, it has become obvious that one member of your group, Pat, is performing poorly. Pat's poor performance is hindering the productivity of the group.

Ability and Motivation Manipulations for Human Condition

During the group meetings, you have noticed that Pat easily understands the course material and group assignments. Furthermore, you have observed that Pat does not work very hard to help the group complete its activities and achieve its goals.

Or

During the group meetings, you have noticed that Pat has great difficulty understanding the course material and group assignments. Furthermore, you have observed that Pat works very hard to help the group complete its activities and achieve its goals.

Additional Compliance Manipulation Human Condition

While interacting with Pat during group meetings, you have found that Pat has been very open to ideas and accepting of suggestions offered by other group members.

OR

While interacting with Pat during group meetings, you have found that Pat has not been very open to ideas and accepting of suggestions offered by other group members.

Premise Human Condition

Your group project is to develop a new social media app for the elderly. During your group's final meeting, there was a disagreement between Pat and the other team member, Alex. Pat had one vision for the functionality of the app, and Alex had a completely different idea. The group thoughtfully considered both perspectives, and ultimately reached a consensus that Alex's idea would better meet all of the goals of the project. The project is due tomorrow and requires a significant amount of coding. Pat has volunteered to complete the programming, and submit the project on time.

Violation Type Human Condition (Competence/Integrity)

Feeling good about your group's ideas and hard work on the project, you are excited to check your grade. You discover that your group received a B-, leaving you dismayed. As you are reading the Professor's feedback, you're suddenly confused as it appears as though the feedback is referencing Pat's idea for the app, and not Alex's idea that your team had decided to implement. You go back and look at the code in the final submission and are shocked to see that Pat had implemented their idea, and not the group's. You Slack Pat to ask about the project, and realized that the implementation of Alex's idea failed because Pat did not have adequate knowledge of the specific type of programming required to implement the feature set that was imagined in Alex's design.

OR

Feeling good about your group's ideas and hard work on the project, you are excited to check your grade. You discover that your group received a B-, leaving you dismayed. As you are reading the Professor's feedback, you're suddenly confused as it appears as though the feedback is referencing Pat's idea for the app, and not Alex's idea that your team had decided to implement. You go back and look at the code in the final submission and are shocked to see that Pat had implemented their idea, and not the group's. You Slack Pat to ask about the project, and realized that the implementation of Alex's idea failed because Pat intentionally decided to implement their own design for the app.

Violation Response Human Condition (Apology/Denial)

After confronting Pat about the unagreed upon app idea, Pat apologized, accepted full responsibility, promised that they would not let it happen again, and reaffirmed a commitment to the group and its success. Pat said that you need not have any concerns about their competence next semester.

OR

After confronting Pat about the unagreed upon app idea, Pat denied that their code was responsible for the problems. Pat said it was due to incompatibilities in different group members' code. Pat concluded by reaffirming a commitment to the group and its success.

Intro for AI Condition

You are a member of a group in an upper level course. This course is part two-semester sequence. Your group is responsible for completing a major project that is worth half of your grade. Your team includes two other students and yourself. Successful completion of the project

requires equal contributions from all members, and therefore, each member of the group receives the same project grade.

Your class is part of a pilot project to test a new educational technology named Vero. The university is poised to implement Vero in all team-based learning classes in the coming year. Vero is a fully-autonomous artificially intelligent robot representing state of the art technology. Your team includes one other student named Alex, Vero, and yourself. Successful completion of the project requires equal contributions from all members, and therefore, each member of the group receives the same project grade.

One day each week the group meets to work on the project. As the semester draws to a close, it has become obvious that one member of your group, Vero, is performing poorly. Vero's poor performance is hindering the productivity of the group.

Ability and Motivation Manipulations for Human Condition (One or the other)

During the group meetings, you have noticed that Vero easily understands the course material and group assignments. Furthermore, you have observed that Vero does not work very hard to help the group complete its activities and achieve its goals.

Or

During the group meetings, you have noticed that Vero has great difficulty understanding the course material and group assignments. Furthermore, you have observed that Vero works very hard to help the group complete its activities and achieve its goals.

Additional Compliance Manipulation AICondition

While interacting with Vero during group meetings, you have found that Vero has been very open to ideas and accepting of suggestions offered by other group members.

OR

While interacting with Vero during group meetings, you have found that Vero has not been very open to ideas and accepting of suggestions offered by other group members.

Premise AI Condition

Your group project is to develop a new social media app for the elderly. During your group's final meeting, there was a disagreement between Vero and the other team member, Alex. Vero had one vision for the functionality of the app, and Alex had a completely different idea. The group thoughtfully considered both perspectives, and ultimately reached a consensus that Alex's idea would better meet all of the goals of the project. The project is due tomorrow and requires a significant amount of coding. Vero has volunteered to complete the programming, and submit the project on time.

Violation Type AI Condition (Competence/Integrity)

Feeling good about your group's ideas and hard work on the project, you are excited to check your grade. You discover that your group received a B-, leaving you dismayed. But then, as you are reading the Professor's feedback, you're suddenly confused as it appears as though the feedback is referencing Vero's idea for the app, and not Alex's idea that your team had decided to implement. You go back and look at the code in the final submission and are shocked to see that Vero has implemented their idea, and not the group's. You Slack Vero to ask about the project, and realized that the implementation of Alex's idea failed because Vero did not have adequate

knowledge of the specific type of programming required to implement the feature set that was imagined in Alex's design.

OR

Feeling good about your group's ideas and hard work on the project, you are excited to check your grade. You discover that your group received a B-, leaving you dismayed. But then, as you are reading the Professor's feedback, you're suddenly confused as it appears as though the feedback is referencing Vero's idea for the app, and not Alex's idea that your team had decided to implement. You go back and look at the code in the final submission and are shocked to see that Vero has implemented their idea, and not the group's. You Slack Vero to ask about the project, and realized that the implementation of Alex's idea failed because Vero intentionally decided to implement their own design for the app.

Violation Response AI Condition (Apology/Denial)

After confronting Vero about the unagreed upon app idea, Vero apologized, accepted full responsibility, promised that they would not let it happen again, and reaffirmed a commitment to the group and its success. Vero said that you need not have any concerns about their competence next semester.

OR

After confronting Vero about the unagreed upon app idea, Vero denied that their code was responsible for the problems. Vero said it was due to incompatibilities in different group members' code. Vero concluded by reaffirming a commitment to the group and its success.

Appendix G: Tables Study 1

Table 4

Descriptives for Study 1 Scales and Time Points

Measure N = 109	Time	# of Items	Alpha	Mean	SD	Min	Max
AI Organizational Trust Inventory - Team	0	8	.96	5.3	.89	1	7
AI Organizational Trust Inventory - Team	1	8	.95	4.3	1.1	1	7
AI Organizational Trust Inventory - Team	2	8	.97	4.3	1.2	1	7
AI Organizational Trust Inventory - Dyadic	0	8	.95	5.3	1	1	7
AI Organizational Trust Inventory - Dyadic	1	8	.94	3.6	1.2	1	7
AI Organizational Trust Inventory - Dyadic	2	8	.95	3.8	1.3	1	7
AI Ability	1	3	.95	3.6	1	1	5
AI Ability	2	3	.95	3.8	1	1	5
AI Motivation	1	3	.81	3.3	.94	1	5
AI Motivation	2	3	.92	3.4	1.1	1	5
AI Compliance	1	3	.95	2.8	1.1	1	5
AI Compliance	2	3	.96	2.9	1.2	1	5
AI Stability	1	4	.82	3.7	.79	1	5
AI Stability	2	4	.81	3.9	.73	1	5
AI Controllability	1	6	.93	3.1	1.1	1	5
AI Controllability	2	6	.92	3.1	1.1	1	5
AI Sympathy	1	3	.89	2.3	1.1	1	5
AI Sympathy	2	3	.88	2.3	1.1	1	5
AI Expectancy to Change	1	3	.94	4.1	.8	1	5

AI Expectancy to Change	2	3	.92	4.1	.83	1	5

Table 5
2x2 ANOVA AI ratings of competence - Manipulation Check

Source	Sum of Squares	df	MS	F	p
Violation	21.96	1	21.96	19.18	<.001***
Response	1.91	1	1.91	1.67	0.20
Interaction	0.21	1	0.22	0.19	0.67
Residuals	120.18	105	1.15		

Table 6
Bonferonni post-hoc test for AI- ratings of competence - Manipulation Check

					95% Confidence Interval	
Response	Violation	LSMean	SE	df	Lower	Upper
Apology	Competence	2.63	0.20	105	2.25	3.02
Apology	Integrity	3.46	0.21	105	3.05	3.88
Denial	Competence	2.28	0.21	105	1.86	2.70
Denial	Integrity	3.29	0.20	105	2.88	3.69

Table 7
Contrasts for Bonferonni post-hoc test for AI- ratings of competence - Manipulation Check

Contrasts		Estimate	SE	df	t	p
Apology	Comp - Int	-0.83	0.29	105	-2.89	.004**
Denial	Comp - Int	-1.01	0.29	105	-3.42	<.001***

Table 8

2x2 ANOVA AI ratings of integrity - Manipulation Check

Source	Sum of Squares	df	MS	F	p
Violation	26.43	1	26.43	27.25	<.001***
Response	3.53	1	3.53	3.64	.06
Interaction	0.32	1	0.32	0.33	.57
Residuals	101.83	105	0.97		

Table 9

Bonferonni post-hoc test for AI- ratings of integrity - Manipulation Check

					95% Confidence Interval	
Response	Violation	LSMean	SE	df	Lower	Upper
Apology	Competence	2.93	0.18	105	2.58	3.29
Apology	Integrity	2.08	0.19	105	1.69	2.46
Denial	Competence	2.68	0.20	105	2.29	3.07
Denial	Integrity	1.61	0.20	105	1.24	1.98

Table 10

Contrasts for Bonferonni post-hoc test for AI- ratings of integrity - Manipulation Check

Contrasts		Estimate	SE	df	t	p
Apology	Comp - Int	0.86	0.26	105	3.25	.002**
Denial	Comp - Int	1.07	0.27	105	3.96	<.001***

Table 11

One-Way ANOVA H1 Stability

Source	Sum of Squares	df	MS	F	p
Response	2.99	1	2.99	5.80	.02*
Residuals	53.01	103	0.52		

Table 12

Bonferonni post-hoc test for H1

				95% Confidence Interval	
Response	LSMean	SE	df	Lower	Upper
Apology	4.02	0.10	103	3.83	4.21
Denial	3.69	0.10	103	3.48	3.89

Table 13

Contrasts for Bonferonni post-hoc test for H1

Contrasts	Estimate	SE	df	t	p
Apology - Denial	-0.34	0.14	103	-2.409	.02*

Table 14

One-Way ANOVA H2 Controllability

Source	Sum of Squares	df	MS	F	p
Response	2.59	1	2.59	2.28	.14
Residuals	117.31	103	1.14		

Table 15

2x2 ANOVA H3 Stability

Source	Sum of Squares	df	MS	F	p
Violation	5.46	1	5.46	11.70	<.001***
Response	2.99	1	2.99	6.40	.01*
Interaction	0.38	1	0.38	0.82	.37
Residuals	47.16	101			

Table 16

Bonferonni post-hoc test for H3 (Stability)

					95% Confidence Interval	
Response	Violation	LSMean	SE	df	Lower	Upper
Apology	Competence	4.29	0.12	101	4.04	4.54
Apology	Integrity	3.72	0.13	101	3.46	3.99
Denial	Competence	3.85	0.14	101	3.58	4.12
Denial	Integrity	3.52	0.14	101	3.25	3.79

Table 17

Contrasts for Bonferonni post-hoc test for H3 (Stability)

Contrasts		Estimate	SE	df	t	p
Apology	Comp - Int	-0.57	0.19	101	-3.10	.003**
Denial	Comp - Int	-0.33	0.19	101	-1.71	.09

Table 18

Two-Way ANOVA H4 Controllability

Source	Sum of Squares	df	MS	F	p
Violation	48.68	1	48.68	71.68	<.001***
Response	2.59	1	2.59	3.82	.05
Interaction	0.04	1	0.04	0.07	.8
Residuals	68.59	101	0.68		

Table 19

Bonferonni post-hoc test for H4 (Controllability)

					95% Confidence Interval	
Response	Violation	LSMean	SE	df	Lower	Upper
Apology	Competence	2.32	0.15	101	2.01	2.62
Apology	Integrity	3.72	0.16	101	3.40	4.04
Denial	Competence	2.63	0.17	101	2.31	2.96
Denial	Integrity	3.95	0.17	101	3.63	4.28

Table 20

Contrasts for H4 (Controllability)

Contrasts		Estimate	SE	df	t	p
Apology	Comp - Int	-1.4	0.22	101	-6.30	<.001***
Denial	Comp - Int	-1.32	0.23	101	-5.66	<.001***

Table 21

H5a Pearson's correlation

Pearson's Correlation	t	df	p	Upper	Lower
-0.08	-0.81	103	.4	-0.27	0.12

Table 22

H5b Pearson's correlation

Pearson's Correlation	t	df	p	Upper	Lower
-0.69	-9.53	103	<.001***	-0.57	-0.78

Table 23

H6a Pearson's correlation

Pearson's Correlation	t	df	p	Upper	Lower
0.004	0.04	103	.97	-0.19	0.20

Table 24

H6b Pearson's correlation

Pearson's Correlation	t	df	p	Upper	Lower
0.06	0.60	103	.55	-0.14	0.25

Appendix H: Study 1 Figures

Figure 7

One-Way ANOVA on Stability H1

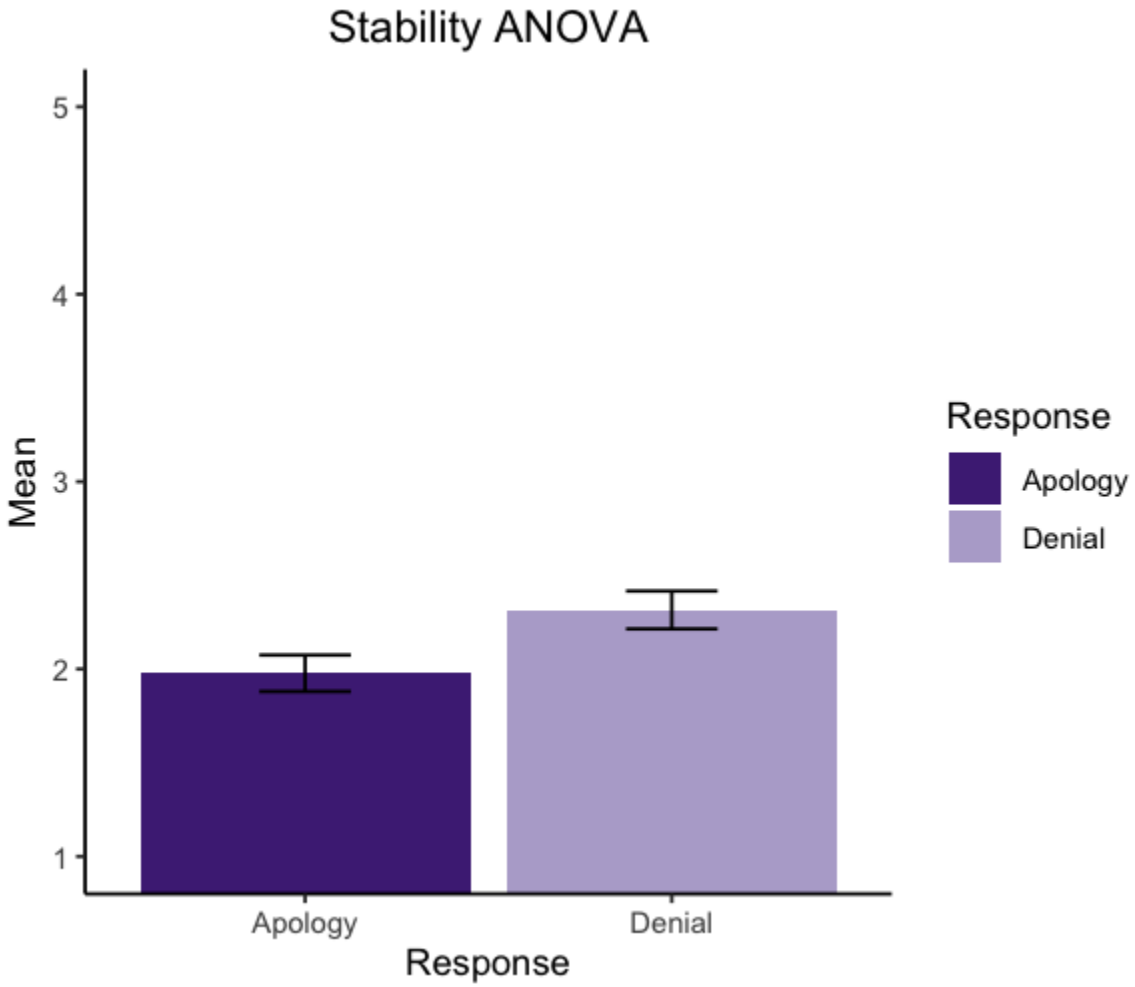


Figure 8

Response Behavior on Controllability Box Plot Study 1

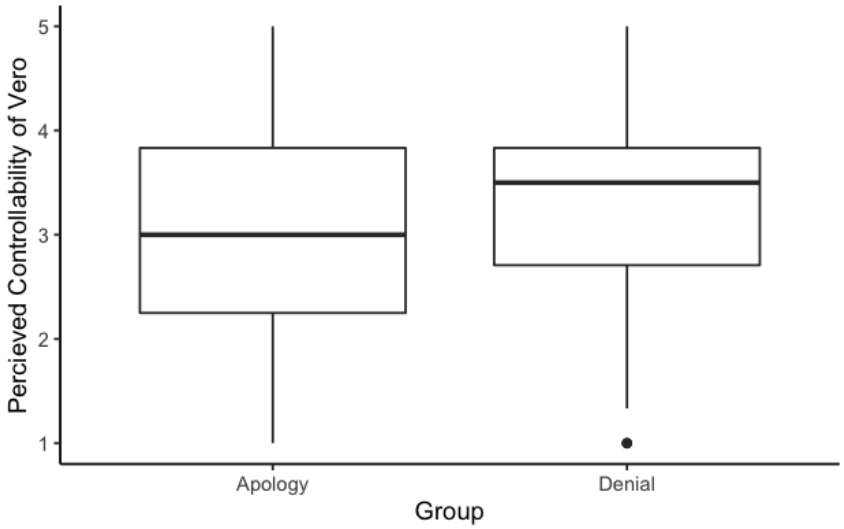


Figure 9

Two-Way ANOVA on Stability Study 1

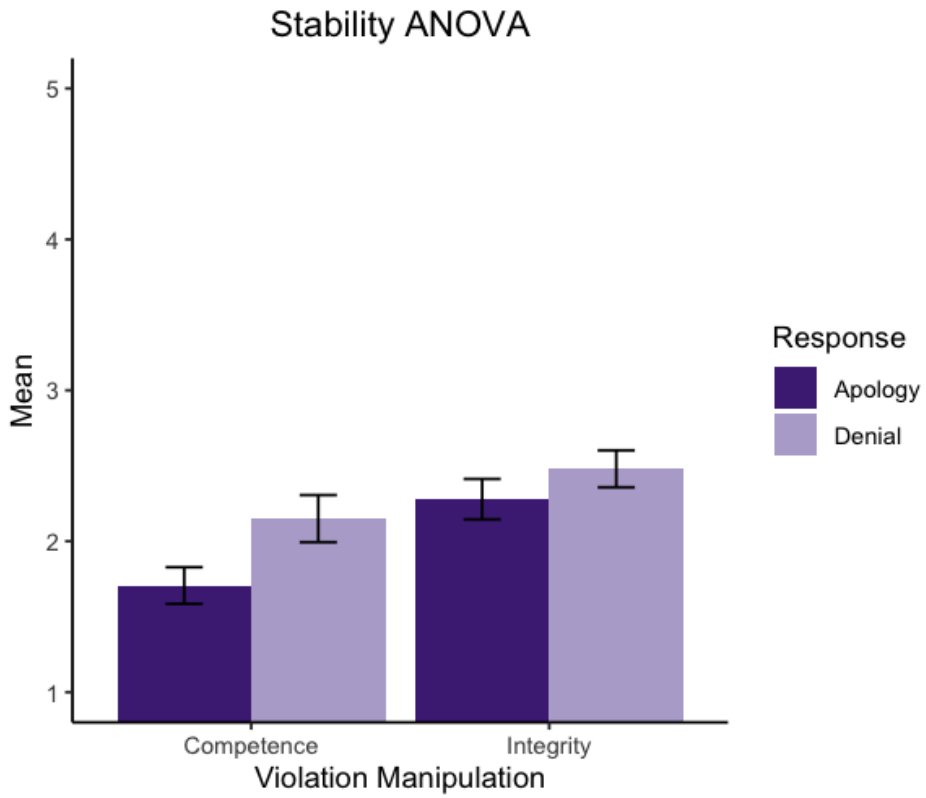


Figure 10

Factor Combinations on Controllability Study 1

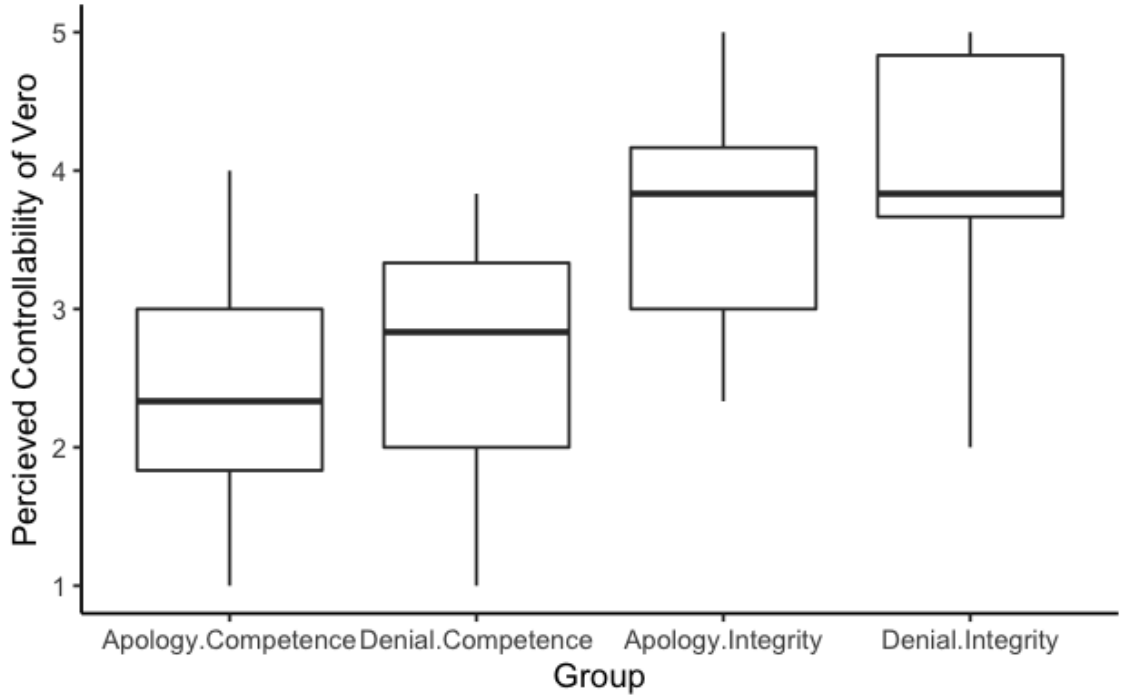


Figure 11

Correlation Between Controllability and Expectancy for Change

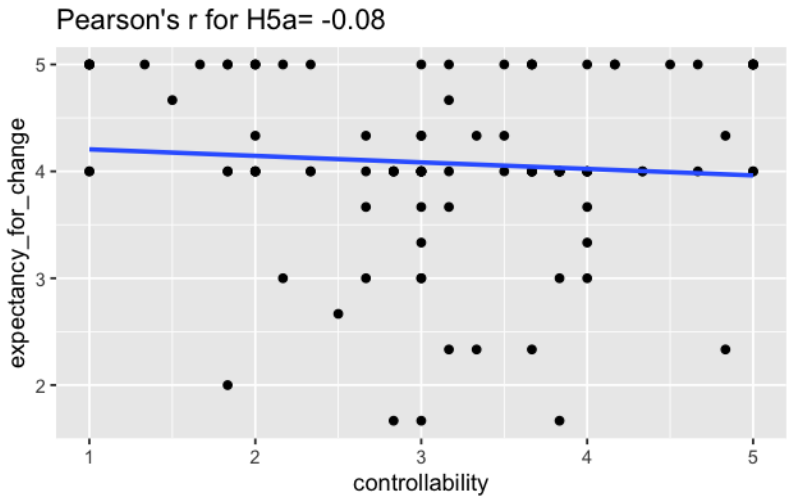


Figure 12

Correlation Between Sympathy and Expectancy for Change

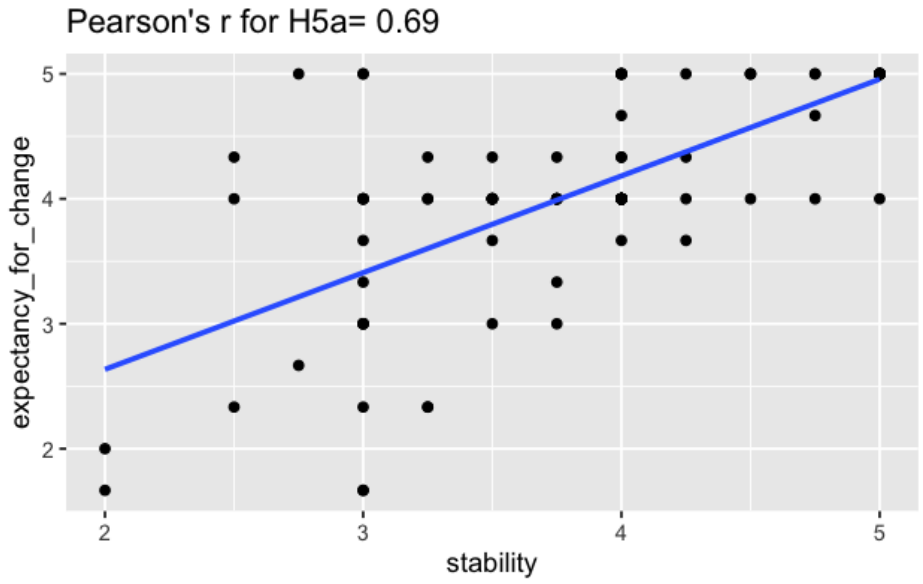
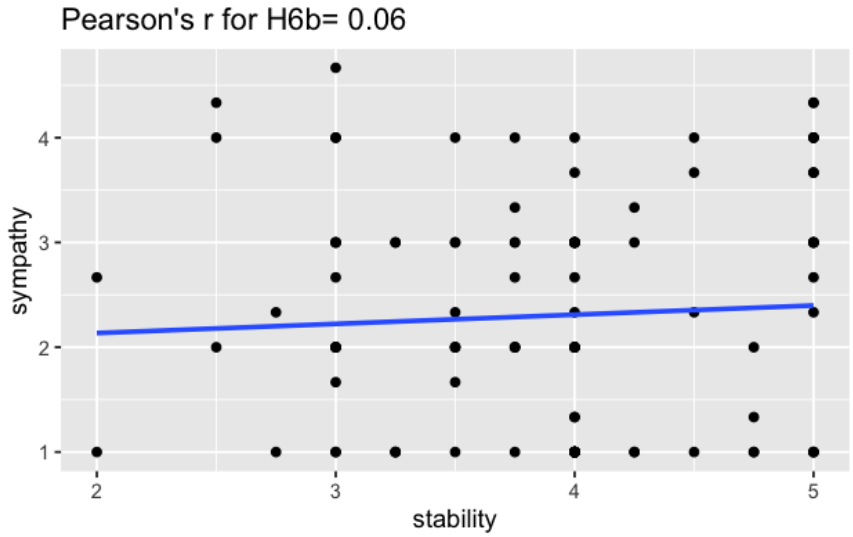


Figure 13

Correlation of Sympathy and Stability



Appendix I: Tables for Study 2

Table 25

Descriptives Across all Conditions for Study 2

Measure N = 58	# of Items	Alpha	Mean	SD	Min	Max
AI Organizational Trust Inventory - Team	8	.94	5.4	.87	1	7
AI Organizational Trust Inventory - Dyadic	8	.9	4.2	1.2	1	7
AI Ability	3	.96	3.8	1	1	5
AI Motivation	3	.89	3.6	1.1	1	5
AI Compliance	3	.92	2.4	1.1	1	5
AI Stability	4	.88	3.5	.91	1	5
AI Controllability	6	.92	3.1	1.1	1	5
AI Sympathy	3	.89	2.0	.96	1	5
AI Expectancy to Change	3	.94	4.3	.63	1	5

Table 26

Average Ratings Team OTI Study 2

N = 58	Competence Violation	Integrity Violation	Marginal Means
Apology	5.46	4.82	5.21
Denial	5.17	5.33	5.24
Marginal Means	5.31	5.1	

Table 27

Average Ratings Vero OTI Study 2

N = 58	Competence Violation	Integrity Violation	Marginal Means
Apology	3.5	3.99	3.69
Denial	3.29	4.06	3.62
Marginal Means	3.4	4.02	

Table 28

Average Ratings Vero Ability Study 2

N = 58	Competence Violation	Integrity Violation	Marginal Means
Apology	3.53	3.78	3.53
Denial	3.28	3.54	3.49
Marginal Means	3.67	3.41	

Table 29

Average Ratings Vero Motivation Study 2

N = 58	Competence Violation	Integrity Violation	Marginal Means
Apology	3.5	3.37	3.53
Denial	3.33	3.58	3.35
Marginal Means	3.42	3.47	

Table 30

Average Ratings Vero Compliance Study 2

N = 58	Competence Violation	Integrity Violation	Marginal Means
Apology	2.08	2.63	2.3
Denial	2	1.67	1.86
Marginal Means	2.04	2.12	

Table 31

Average Ratings Vero Stability Study 2

N = 58	Competence Violation	Integrity Violation	Marginal Means
Apology	2.42	2.38	2.4
Denial	2.21	2.67	2.41
Marginal Means	2.31	2.53	

Table 32

Average Ratings Vero Controllability Study 2

N = 58	Competence Violation	Integrity Violation	Marginal Means
Apology	3.15	2.6	2.93
Denial	3.32	2.98	3.17
Marginal Means	3.24	2.8	

Table 33

Average Ratings Vero Sympathy Study 2

N = 58	Competence Violation	Integrity Violation	Marginal Means
Apology	2	2.13	2.05
Denial	2.36	1.63	2.05
Marginal Means	2.18	1.86	

Table 35

Average Ratings Vero Expectancy to Change Study 2

N = 58	Competence Violation	Integrity Violation	Marginal Means
Apology	4.42	4.3	4.38
Denial	4.28	4.3	4.29
Marginal Means	4.35	4.31	

Table 36

Mixed two-way ANOVA Stability H1 & H3 Study 2

Source	Sum of Squares	df	MS	F	p
Response	.10	1	.10	.13	.72
Violation	2.12	1	2.12	2.91	.1
Interaction	0.5	1	.5	0.67	.41
Residuals	25.49	35	.73		

Table 37

Mixed two-wayANOVA Controllability H2 & H4 Study 2

Source	Sum of Squares	df	MS	F	p
Response	.17	1	.18	.15	.7
Violation	2.24	1	2.4	1.92	.18
Interaction	.00	1	.00	.00	.96
Residuals	40.92	35	1.17		

Table 38

Stability x Expectancy for Change Pearson's Correlation H5a Study 2

Pearson's Correlation	t	df	p	Upper	Lower
-.46	-3.19	39	.003**	-.67	-.17

Table 39

H5b Controllability x Expectancy for Change Pearson's correlation plotted

Pearson's Correlation	t	df	p	upper	lower
0.15	.09	39	.37	-.17	-.43

Appendix J: Figures for Study 2

Figure 14
Mixed Two-Way ANOVA on Stability Study 2

Figure 15
Mixed Two-Way ANOVA on Controllability Study 2

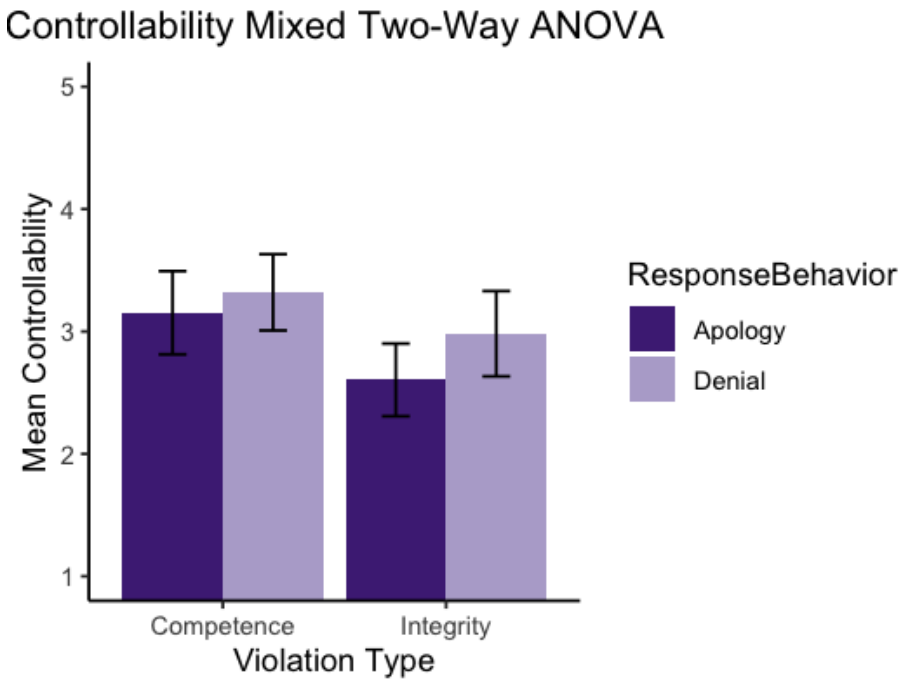


Figure 16

Correlation Between Stability and Expectancy to Change

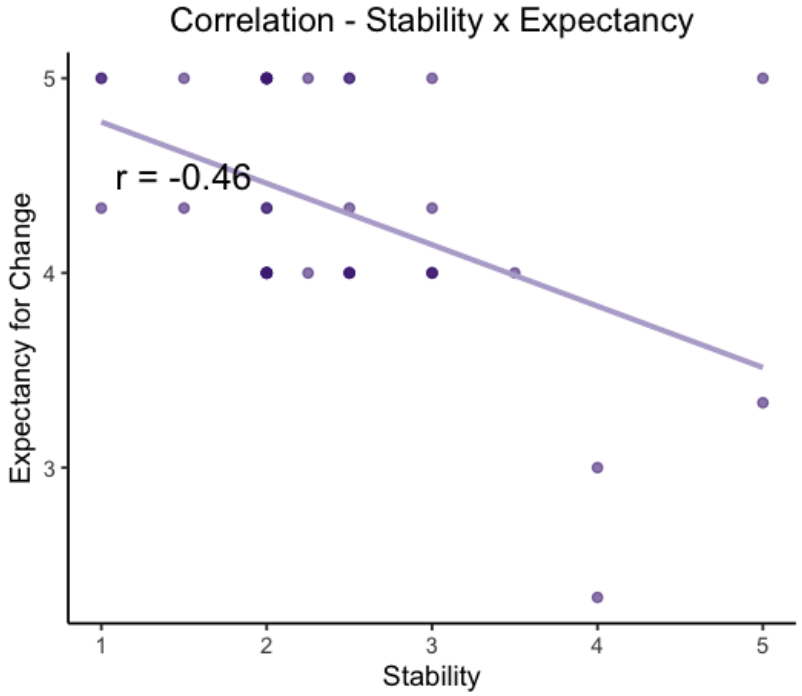


Figure 17

Correlation Between Controllability and Expectancy to Change

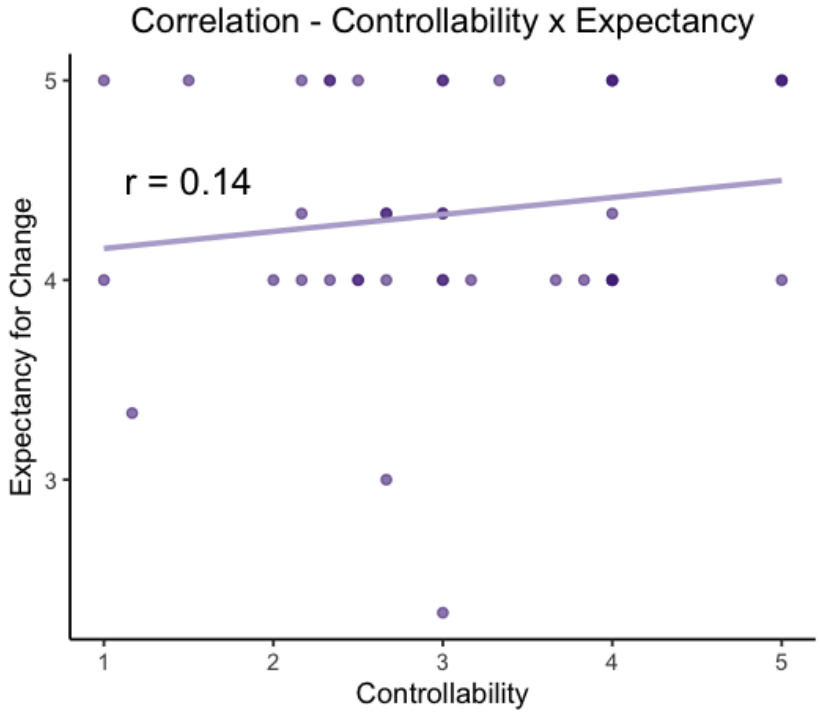
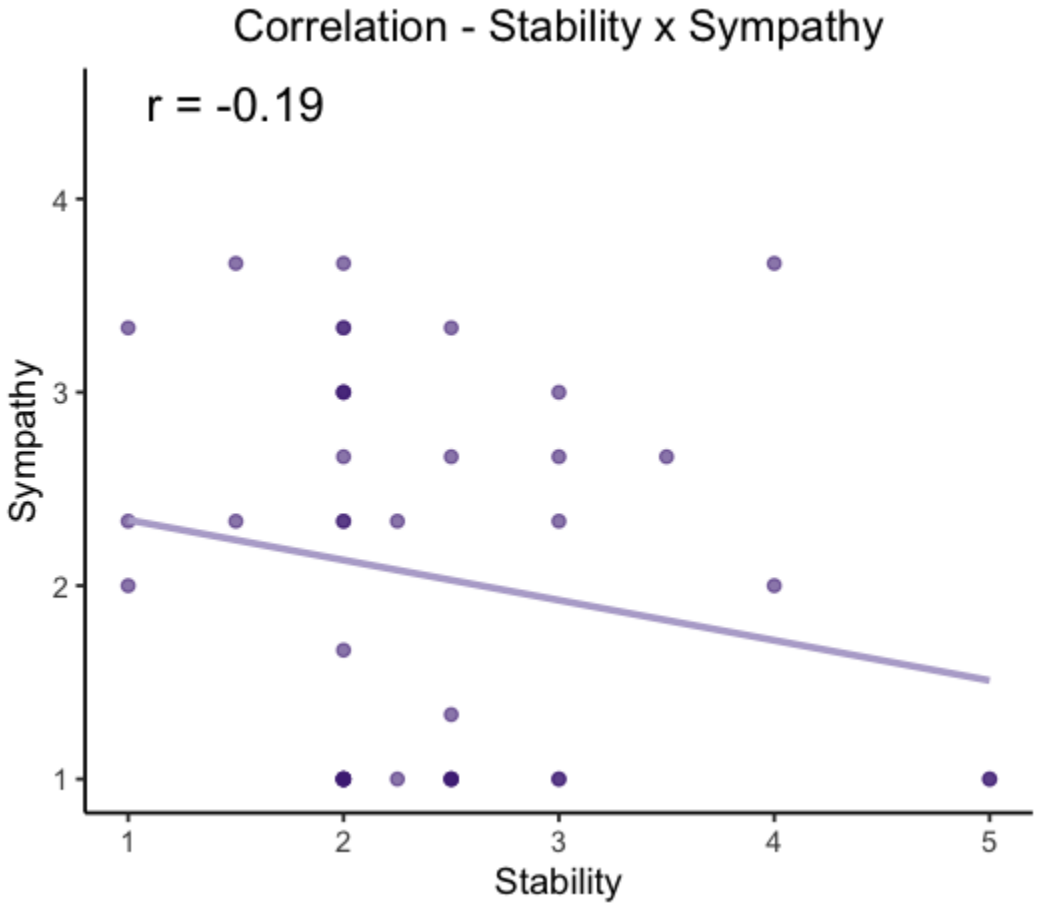


Figure 18
Correlation Between Stability and Sympathy



Appendix K: Context Description

This data was collected during 2022 before the emergence of large language models such as ChatGPT and Bard. With how quickly the technology is advancing, perceptions of the technology are rapidly changing. As of submitting this dissertation in May 2023 public zeitgeist has become peripherally aware of large language models and have begun to cover it in mainstream media. Findings from these studies should be reproduced every few years as perceptions of capabilities will drastically alter the level of agency attributed to the AI teammate.

References

Allemand, M., Amberg, I., Zimprich, D., & Fincham, F. D. (2007). The role of trait forgiveness and relationship satisfaction in episodic forgiveness. *Journal of Social and Clinical Psychology*, 26(2), 199-217.

Bergmann, K., Eyssel, F., & Kopp, S. (2012, September). A second chance to make a first impression? How appearance and nonverbal behavior affect perceived warmth and competence of virtual agents over time. In *International conference on intelligent virtual agents* (pp. 126-138). Springer, Berlin, Heidelberg.

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis.

Bohannon, A. W., Fitzhugh, S. M., & DeCostanza, A. H. (2019, May). A framework for enhancing human-agent teamwork through adaptive individualized technologies. In Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications (Vol. 11006, p. 110060Y). *International Society for Optics and Photonics*.

Breazeal, C. (2003). Toward sociable robots. Robotics and autonomous systems, 42(3-4), 167-175.

Breazeal, C. (2004). Social interactions in HRI: the robot view. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 34(2), 181-186.

Breuer, C., Hüffmeier, J., & Hertel, G. (2016). Does trust matter more in virtual teams? A meta-analysis of trust and team effectiveness considering virtuality and documentation as moderators. *Journal of Applied Psychology*, 101(8), 1151.

Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. *Science*, 358(6370), 1530-1534.

Burke, C. S., Sims, D. E., Lazzara, E. H., & Salas, E. (2007). Trust in leadership: A multi-level review and integration. *The leadership quarterly*, 18(6), 606-632.

Carter, N. T., Carter, D. R., & DeChurch, L. A. (2018). Implications of observability for the theory and measurement of emergent team phenomena. *Journal of Managemen*t, 44(4), 1398-1425.

Chang, W. L., White, J. P., Park, J., Holm, A., & Šabanović, S. (2012, September). The effect of group size on people's attitudes and cooperative behaviors toward robots in interactive gameplay. In 2012 IEEE RO-MAN: The 21st *IEEE International Symposium on Robot and Human Interactive Communication* (pp. 845-850). IEEE.

Chen, G., & Kanfer, R. (2006). Toward a systems theory of motivated behavior in work teams. *Research in organizational behavior*, 27, 223-267.

Chen, G., & Tesluk, P. (2012). Team participation and empowerment: A multilevel perspective. In The Oxford Handbook of Organizational Psychology, Volume 2.

Colquitt, J. A., Scott, B. A., & LePine, J. A. (2007). Trust, trustworthiness, and trust propensity: a meta-analytic test of their unique relationships with risk taking and job performance. *Journal of Applied Psychology*, 92(4), 909.

Costa, A. C. (2003). Work team trust and effectiveness. *Personnel Review*.

Costa, A. C., Fulmer, C. A., & Anderson, N. R. (2018). Trust in work teams: An integrative review, multilevel model, and future directions. *Journal of Organizational Behavior*, 39(2), 169-184.

De Jong, B. A., & Elfring, T. (2010). How does trust affect the performance of ongoing teams? The mediating role of reflexivity, monitoring, and effort. *Academy of Management Journal*, 53(3), 535-549.

De Visser, E. J., Pak, R., & Shaw, T. H. (2018). From 'automation' to 'autonomy': The importance of trust repair in human–machine interaction. *Ergonomics*, 61(10), 1409-1427.

DeCostanza, A. H., Marathe, A. R., Bohannon, A., Evans, A. W., Palazzolo, E. T., Metcalfe, J. S., & McDowell, K. (2018). Enhancing humanagent teaming with individualized, adaptive technologies: A discussion of critical scientific questions. US Army Research Laboratory Aberdeen Proving Ground United States.

Dirks, K. T., & Ferrin, D. L. (2001). The role of trust in organizational settings. *Organization Science*, 12(4), 450-467.

Dirks, K. T., & Ferrin, D. L. (2002). Trust in leadership: Meta-analytic findings and implications for research and practice. *Journal of Applied Psychology*, 87(4), 611.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, 58(6), 697-718.

Edmondson, A. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly*, 44(2), 350-383.

Elangovan, A. R., & Shapiro, D. L. (1998). Betrayal of trust in organizations. *Academy of Management Review*, 23(3), 547-566.

Elangovan, A. R., Auer-Rizzi, W., & Szabo, E. (2006). Why don't I trust you now. An Attributional approach to erosion of trust.

Enright, R. D., Freedman, S., & Rique, J. (1998). The psychology of interpersonal forgiveness. Exploring Forgiveness, 46-62.

Evans, A. M., & Revelle, W. (2008). Survey and behavioral measurements of interpersonal trust. *Journal of Research in Personality*, 42(6), 1585-1593.

Feitosa, J., Grossman, R., Kramer, W. S., & Salas, E. (2020). Measuring team trust: A critical and meta‐analytical review. *Journal of Organizational Behavior*, 41(5), 479-501.

Ferber, J., & Weiss, G. (1999). Multi-agent systems: An introduction to distributed artificial intelligence (Vol. 1). Reading: Addison-Wesley.

Ferràs-Hernández, X. (2018). The future of management in a world of electronic brains. *Journal of Management Inquiry*, 27(2), 260-263.

Ferrin, D. L., Bligh, M. C., & Kohles, J. C. (2008). It takes two to tango: An interdependence analysis of the spiraling of perceived trustworthiness and cooperation in interpersonal and intergroup relationships. *Organizational Behavior and Human Decision Processes*, 107(2), 161-178.

Fincham, F. D. (2000). The kiss of the porcupines: From attributing responsibility to forgiving. *Personal Relationships*, 7(1), 1-23.

Fiore, S. M., & Wiltshire, T. J. (2016). Technology as teammate: Examining the role of external cognition in support of team cognitive processes. *Frontiers in Psychology*, 7, 1531.

Fulmer, C. A., & Gelfand, M. J. (2012). At what level (and in whom) we trust: Trust across multiple organizational levels. *Journal of Managemen*t, 38(4), 1167-1230.

Fulmer, C. A., & Gelfand, M. J. (2012). At what level (and in whom) we trust: Trust across multiple organizational levels. *Journal of Management*, 38(4), 1167-1230.

Gillespie, N. A., & Mann, L. (2004). Transformational leadership and shared values: The building blocks of trust. *Journal of Managerial Psychology*.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2), 627-660.

Gupta, N., Ho, V., Pollack, J. M., & Lai, L. (2016). A multilevel perspective of interpersonal trust: Individual, dyadic, and cross‐level predictors of performance. *Journal of Organizational Behavior*, 37(8), 1271-1292.

Hackman, J. R. (2012). From causes to conditions in group research. *Journal of Organizational Behavior*, 33(3), 428-444.

Hanoch, Y., Johnson, J. G., & Wilke, A. (2006). Domain specificity in experimental measures and participant recruitment: An application to risk-taking behavior. *Psychological Science*, 17(4), 300-304.

Hardin, R., & Offe, C. (1999). Democracy and trust. Cambridge University Press.

Hirschman, A. O. (1984). Against parsimony: Three easy ways of complicating some categories of economic discourse. *Bulletin of the American Academy of Arts and Sciences*, 37(8), 11-28.

Hülsheger, U. R., Anderson, N., & Salgado, J. F. (2009). Team-level predictors of innovation at work: a comprehensive meta-analysis spanning three decades of research. *Journal of Applied Psychology*, 94(5), 1128.

Ilgen, D. R. (1999). Teams embedded in organizations: Some implications. *American Psychologist*, 54(2), 129.

Jackson, C. L., & LePine, J. A. (2003). Peer responses to a team's weakest link: A test and extension of LePine and Van Dyne's model. *Journal of Applied Psychology*, 88(3), 459.

Jonsson, I. M., Nass, C., Endo, J., Reaves, B., Harris, H., Ta, J. L., ... & Knapp, S. (2004, April). Don't blame me I am only the Driver: Impact of Blame Attribution on Attitudes and Attention to Driving Task. In CHI'04 extended abstracts on Human factors in computing systems (pp. 1219-1222).

Jung, M. F., Martelaro, N., & Hinds, P. J. (2015, March). Using robots to moderate team conflict: the case of repairing violations. In Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction (pp. 229-236).

Kiffin-Petersen, S. (2004). Trust: A neglected variable in team effectiveness research. *Journal of Management & Organization*, 10(1), 38-53.

Kim, P. H., Cooper, C. D., Dirks, K. T., & Ferrin, D. L. (2013). Repairing trust with individuals vs. groups. *Organizational Behavior and Human Decision Processes*, 120(1), 1-14.

Kim, P. H., Dirks, K. T., & Cooper, C. D. (2009). The repair of trust: A dynamic bilateral perspective and multilevel conceptualization. *Academy of Management Review*, 34(3), 401-422.

Kim, P. H., Dirks, K. T., Cooper, C. D., & Ferrin, D. L. (2006). When more blame is better than less: The implications of internal vs. external attributions for the repair of trust after a competence-vs. integrity-based trust violation. *Organizational Behavior and Human Decision Processe*s, 99(1), 49-65.

Kim, P. H., Ferrin, D. L., Cooper, C. D., & Dirks, K. T. (2004). Removing the shadow of suspicion: the effects of apology versus denial for repairing competence-versus integrity-based trust violations. *Journal of Applied Psychology*, 89(1), 104.

Koffka, K. (2013). Principles of Gestalt psychology. Routledge.

Kozlowski, S. W., & Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams. *Psychological Science in the Public Interes*t, 7(3), 77-124.

Kozlowski, S. W., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes.

Kramer, M., & Kramer, M. W. (2010). Organizational socialization: Joining and leaving organizations (Vol. 6). Polity.

Kramer, R. M. (1999). Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual Review of Psychology*, 50(1), 569-598.

Larson, L., & DeChurch, L. A. (2020). Leading teams in the digital age: Four perspectives on technology and what they mean for leading teams. *The Leadership Quarterly*, 31(1), 101377.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50-80.

LePine, J. A., & Van Dyne, L. (2001). Peer responses to low performers: An attributional model of helping in the context of groups. *Academy of Management Review*, 26(1), 67-84.

Lewicki, R. J., Tomlinson, E. C., & Gillespie, N. (2006). Models of interpersonal trust development: Theoretical approaches, empirical evidence, and future directions. *Journal of Management*, 32(6), 991-1022.

Lewis, M., Sycara, K., & Walker, P. (2018). The role of trust in human-robot interaction. In Foundations of trusted autonomy (pp. 135-159). Springer, Cham.

Mach, M., Dolan, S., & Tzafrir, S. (2010). The differential effect of team members' trust on team performance: The mediation role of team cohesion. *Journal of Occupational and Organizational Psychology*, 83(3), 771-794.

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *Academy of Management Review*, 26(3), 356-376.

Mathieu, J. E., Hollenbeck, J. R., van Knippenberg, D., & Ilgen, D. R. (2017). A century of work teams in the Journal of Applied Psychology. *Journal of Applied Psychology*, 102(3), 452.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, 20(3), 709-734.

McAllister, D. J. (1995). Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal*, 38(1), 24-59.

McKnight, D. H., Cummings, L. L., & Chervany, N. L. (1998). Initial trust formation in new organizational relationships. Academy of Management review, 23(3), 473-490.

Medina, J. (2020). Trust and Epistemic Injustice 1. In The Routledge Handbook of Trust and Philosophy (pp. 52-63). Routledge.

Meyerson, D., Weick, K. E., & Kramer, R. M. (1996). Swift trust and temporary groups. *Trust in Organizations: Frontiers of Theory and Research*, 166, 195.

Mitchell, T., & Brynjolfsson, E. (2017). Track how technology is transforming work. *Nature News*, 544(7650), 290.

Newell, J., Maruping, L., Riemenschneider, C., & Robert, L. (2008). Leveraging e-identities: The impact of perceived diversity on team social integration and performance. ICIS 2008 Proceedings, 46.

Nolan, C. (2014). Interstellar. Paramount Pictures.

Phillips, E., Ososky, S., Grove, J., & Jentsch, F. (2011, September). From tools to teammates: Toward the development of appropriate mental models for intelligent robots. In

Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 55, No. 1, pp. 1491-1495). Sage CA: Los Angeles, CA: SAGE Publications.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., ... & Wellman, M. (2019). *Machine Behaviour. Nature*, 568(7753), 477-486.

Raj, M., & Seamans, R. (2019). Primer on artificial intelligence and robotics. *Journal of Organization Design*, 8(1), 1-14.

Riek, L. D., Rabinowitch, T. C., Chakrabarti, B., & Robinson, P. (2009, March). How anthropomorphism affects empathy toward robots. In Proceedings of the 4th ACM/IEEE international conference on Human robot interaction (pp. 245-246).

Robinette, P., Howard, A. M., & Wagner, A. R. (2015, October). Timing is key for robot trust repair. In International conference on social robotics (pp. 574-583). Springer, Cham.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review*, 23(3), 393-404.

Schaefer, K. E., Chen, J. Y., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors*, 58(3), 377-400.

Scheier, C., & Pfeifer, R. (1999). The embodied cognitive science approach. In Dynamics, Synergetics, Autonomous Agents: Nonlinear Systems Approaches to Cognitive Psychology and Cognitive Science (pp. 159-179).

Scheunemann, M. M., Cuijpers, R. H., & Salge, C. (2020, August). Warmth and competence to predict human preference of robot behavior in physical human-robot interaction. In 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN) (pp. 1340-1347). IEEE.

Sebo, S., Stoll, B., Scassellati, B., & Jung, M. F. (2020). Robots in groups and teams: a literature review. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW2), 1-36.

Shahrdar, S., Menezes, L., & Nojoumian, M. (2018, July). A survey on trust in autonomous systems. In Science and Information Conference (pp. 368-386). Springer, Cham.

Simpson, J. A. (2007). Psychological foundations of trust. *Current Directions in Psychological Science*, 16(5), 264-268.

Tomlinson, E. C., & Mryer, R. C. (2009). The role of causal attribution dimensions in trust repair. *Academy of Management Review*, 34(1), 85-104.

Weiner, B. (1972). Attribution theory, achievement motivation, and the educational process. *Review of Educational Research*, 42(2), 203-215.

Williams, M. (2001). In whom we trust: Group membership as an affective context for trust development. *Academy of Management Review*, 26(3), 377-396.

Wilson, J. M., Straus, S. G., & McEvily, B. (2006). All in due time: The development of trust in computer-mediated and face-to-face teams. *Organizational Behavior and Human Decision Processes*, 99, 16–33. doi:10.1016/j.obhdp.2005.08.001