Cognitive Computing and Internet of Things, Vol. 73, 2023, 169–178
https://doi.org/10.54941/ahfe1003294

Trust: The Vital Fluid of Interactions

Laurent M. Chaudron[1,2], Jean-Marie Burkhardt[3], Lisa Chouchane[4], Pauline Muñoz[2], Nicolas Maille[5], and Anne-Lise Marchand[2]

[1]Theorik-Lab, Cognitive Sciences Institute, France
[2]École de l'Air et de l'Espace Research Center (CREA), France
[3]Gustave Eiffel University, France
[4]French Air Force, France
[5]Onera, Salon de Provence, France

ABSTRACT

In this paper we present a multidimensional approach to the concept of Trust in complex technical environments. A survey allows to focus on the two main types: Interpersonal Trust (IT) and Trust in Automation (TA), an extended generic conceptual modeling based on a grounded theory methodology is proposed; finally the first phase of an experimental campaign is presented. The whole project aims at a better understanding of the role of the various components of Trust during the decision process of a human operator in cooperation with sophisticated systems and human partners.

Keywords: Trust, Human operators, Trustor, Trustee, Automaton, AI, Decision

INTRODUCTION

The rise of all kinds of automatism and their mixture in our daily lives challenges our consciences in a multifaceted way; the unconscious matter that constituted the Trust concept becomes palpable and brings out explicit questions. The aim of this paper is to help to understand the Trust mechanism among operators of complex sociotechnical systems thanks to a dual theoretical-experimental approach.

Technological advances nowadays allow drivers (of cars or trains) or pilots to delegate all or part of the driving: automatic pilots, driverless shuttles, autonomous cars… More generally, from sociology to economics through cybernetics, the concept of trust appears as a pivotal ingredient of Homo Sapiens as a social agent. However, the definitions of trust depend strongly on the field and many studies focus either on a specific approach (sociology, cognitive psychology, ...) or on the applied fields of trust (management, cyber-security, ...). The academic literature has experienced a very significant acceleration in recent years: Trust gets about 5 million entries on scientific search engines. Hence, many questions arise both for highly trained operators of aerospace vehicles to our everyday life regarding the decision process during the management of a mission or a simple mere action: should I trust these AI based systems enjoining me to choose this option, or would it be safer to trust my teammate or my friend telling me to make another choice? And what does trust mean precisely in such a context?

Methodology

A multidisciplinary research team decided to tackle this scientific challenge with three main components:

- ergonomics and cognitive psychology: wide review and the design of a pluri-annual experimental campaign,
- engineering approach: through a grounded theory approach, qualitative analyses of operational cases and simulations,
- symbolic artificial intelligence and knowledge representation: definition of a conceptual generic formal model.

The reader will easily recognize the different categories of contributions among the paper which is organized as follows: - after a literature review from which emerged invariants including a common concept joining IT and TA, - the principles of Trust as a concept are stated so as to build a generic trust model, TorTeeX, - then an experimental campaign is presented including the use of the PAS grid in a first experiment based on a decision task in an aeronautical context, - discussion and some major perspectives are finally proposed.

THE LITERATURE ON TRUST

Among the huge amount of publications related to trust, few studies yet propose to compare trust in automation (TA) and interpersonal trust (IT). The study presented here is part of a larger program aiming at understanding the mechanism of trust within complex sociotechnical systems through an approach combining modeling and experimentation. In this first step, the TorTeeX model is presented, followed by a first experimental study. In the context of an individual's arbitration in favor of one or the other partner, this study aims at understanding to what extent elements related to the trustor's personality can modulate the way he/she trusts a human partner and an automaton partner. In a correlated way, this study explores whether this influence is the same whatever the nature of the trustee.

Because it influences behavior in terms of performance, workload, situational awareness, monitoring, or command recovery (Lee & Moray, 1994; Riley, 1994, Lee & See, 2004, Dirks & Ferrin, 2002; Colquitt, Scott, & LePine, 2007; Breuer, Hüffmeier, & Hertel, 2016; De Jong, Dirks, & Gillespie, 2016; Bollon, 2019; 2020), trust is central to models of expert cognition. In addition to Kahneman & Klein (2009), this central role is also identifiable in Endsley's Situational Awareness model. The latter (2017) mentions a "confidence level" that allows one to associate a degree of uncertainty with the mental model, and to project oneself into future situations. However, she does not develop this mechanism, nor the stage of the process in which it intervenes; she does not explicitly speak of trust or reliability. Similarly, the models used in the literature on trust in automated systems are based on the work of Muir (Muir, 1987; Muir & Moray, 1996), who implies that trust in automated systems is similar to the trust studied in psychology by Deutsch (1973), Rotter (1980) and Rempel et al. However, Muir does not detail or compare these different mechanisms. In other words, the dominant theoretical models of expert activity in complex and dynamic sociotechnical systems (Kahneman & Klein, 2009; Endsley, 1995 or 2017) recognize the essential role of trust for performance but are sometimes evasive about its fundamental nature.

Thus, although these processes draw on the same initial literature and can be described as a trustor's assessment of a trustee's trustworthiness in order to assign him or her a given task in a given context, little literature examines what distinguishes them and what brings them together. Interpersonal trust (IT) assesses the reliability of a third party, which brings it closer to trust in Automation (TA). The theoretical models of IT and TA are close; for example, TA is defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee and See, 2004; p. 54) and IT as "a willingness to accept vulnerability" (Mayer and Davis & Schoorman, 1995). Hoff & Bashir distinguish them (2015), however, based on the trustor's expectations: the trustee's skill, integrity, and benevolence for IT (Mayer et al., 1995), and the machine's usefulness, predictability, and intent for TA (Lee & See, 2007). However, this distinction is perhaps due to the methodologies and theoretical fields mobilized on these two variants of trust; on the one hand, the field of IT is studied mainly through field studies and subjective data from questionnaires (Dirks & Ferrin, 2002) in management sciences; on the other hand, the field of "trust in automation" is based on behavioral data and experimental methodologies in psychology (Hoff & Bashir, 2015). The literature thus offers few experimental benchmarks for understanding and manipulating IT.

To sum up, the current literature on the different types of trust does consider the link between reliability and trust, and even associates trust with the evaluation of the trustee's reliability. However, they do not explore the purpose(s) of these evaluations. Yet, the common cognitive principles stated by Kahneman & Klein or Endsley's model identify the key role of trust in decision-making: it therefore seems essential to understand how the individual uses trust in decision-making. These theoretical comparisons of the different conceptualizations of trust (IT vs. TA) allow postulating that these different forms of trust are in fact variations of the same process. One objective of this study is therefore to contribute to this understanding through a common model of trusts (hereafter), tested through a first experimentation (next section). More precisely, this experimentation aims at evaluating to what extent elements related to the personality of trustor can modulate the way they trust a human partner (IT) or an automaton (TA) in a context of arbitration between the two trustees. In a correlated manner, we ask whether this influence is the same regardless of the nature of the trustee.

A GENERIC MODEL OF TRUST

A study (Chouchane 2022) was conducted following a grounded theory methodology: - qualitative analysis of 100 papers and 60 events (aeronautical, railways, traffic, space, maritime …), - interviews of 6 safety experts from the 3 domains (air & space, ground, sea), - on-line experiment with

80 volunteers, - one mathlab simulator, and during these analyses an inductive model was constructed using classical AI knowledge representation techniques. The synthetic definition of Trust in this model is: *one agent (generally a human-being), called the* ***Trustor****, relies on other agents, the* ***Trustees****, about a subject* ***X*** *(action or mission…) with specific expectations*. The Trustor is the agent who, facing a decision related to X, is in a position to place his trust in other agents: "her·his" Trustees, who symmetrically must inspire her·him with confidence. A mnemonics is: *Tor trusts Tees about X*.

Three pragmatic principles emerged through this study:

1) Trust is an agent-centered concept: the process runs within the Trustor's cognition.
2) Trust consciously emerges (in the Trustor's cognition) during the decision process when a choice must be made between different or dissonant data provided by the Trustees, involving uncertainty, risk and time pressure.
3) Trust is a meta-cognitive process within the Trustor's mind. Trust does not concern the processing of information at a standard level but rather a reflection that the Trustor leads on her·his own processing so as to guide the decision.

Thanks to these 3 principles, a prototypical situation of trust has been aggregated: the Trustor formulates to his various Trustees various requests (explicit or implicit; sometimes negotiable) for information or actions about X. The responses of the Trustee.s provide the Trustor with values of dissonance/consistency and uncertainty/predictability. The Generic Trust Model (Chaudron & al. 2022) called **TorTeeX**, was designed taking into account three categories of trustees encountered among the operational events: - Automatons, - AI based systems, - Human beings:

The TorTeeX model also captures the concept of self-trust, i.e. when Trustee=Trustor, as well as the variability of *sources* or *effectors* as Trustees. The implementation of the TorTeeX model in a multi-agent formal model and the development of the mathematical groundings are currently under study and will not be detailed here. The analysis of the differences between an AI based system and an automaton is the next step of the study.

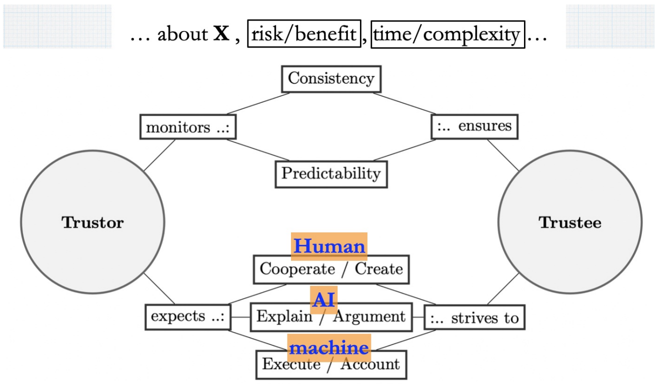

Figure 1: The TorTeeX model.

In the sequel, the prototypical situation studied is a particular case of the TorTeeX model: the Trustor, in order to fulfill her.his mission, has to make a choice between two different pieces of information provided by two Trustees: - an automaton, - a human teammate. The trust in each Trustee will be suggested through a reliability value. The simulated operational situation X gets a high risk score value, and the decision is time constrained. Thus, the meta-cognition of the role of trust in the Trustor mind is activated and explicitly measured thanks to a scale quotation.

EXPERIMENTAL CAMPAIGN

A first step consisted in observing the characteristics of the trustor likely to influence the preference he would have to favor or not an information given by the automatisation rather than an information given by a human partner (Colquitt et al., 2007; Hancock et al., 2011). One of the characteristics that seem to influence human attitudes towards people or automation are individual differences. Individual differences are considered to be constructs such as personality, bias and preference (Lyons & Guznov, 2019). The trust literature presents several individual differences related to trust such as expectations about the trustee (Dzindolet et al., 2002), the personality of the trustor (Merrit & Ilgen, 2008), the genetics of the trustor (Parasuraman et al., 2012) or the working memory of the trustor (Rovira et al., 2017). These studies show that trust is related to some individual differences.

The study presented here focuses on individual measures and more specifically on the trustor's expectations of trustees. According to Mandhavan and Wiegmann's (2007) model, expectations of a machine are different from expectations of a human. The trustor expects imperfections when collaborating with a human trustee, because humans are fallible. However, when the trustor collaborates with an automated agent, he expects the automated agent to perform almost perfectly. As expectations differ between trustees, the trustor adopts a different behavior depending on the nature of the agent. Indeed, the study by Dzindolet et al. (2002) shows that humans are more sensitive to the errors of an automated system than to the errors of another human being. When a human trustee makes a mistake, the trustor takes less account of the human error than the error of the automation. This behavior is explained by the fact that the trustor mentalizes the human trustee as fallible and appeals to a 'human imperfection schema' (Madhavan & Wiegma, 2007). In the case of automated trustees, the human uses a cognitive schema in which the human overestimates the performance of an automated aid, and perceives the automations as 'perfect'. Therefore, when the automation makes a mistake, this "perfect automation scheme" is broken. As a result, the trustor is surprised because he/she had not anticipated this error, and dwells on it. Dzindolet et al. (2002), explain this awareness of the error of automation by the existence of a "perfect automation schema" more or less developed according to individuals

Following this finding, Merritt et al. (2015) develop a measure capable of assessing a person's level of perfect automation schema. A high level of perfect automation schema (PAS) would be characterized by the belief that a technology is infallible. In contrast, people with low levels of automation schema view technology with a skeptical eye (Lyons et al., 2017). Several experimental studies show that the PAS score is related to confidence in automation (Lyons & Guznov, 2019; Merritt et al., 2015). The findings of both studies show that a high PAS is related to a high level of confidence in an automaton. The study presented here builds on and is inspired by the work of Lyons & Guznov (2019) and Merritt et al. The first objective of the presented experimentation, is to confirm that the experimentation conducted is valid and confirms the results of the colleagues. The second objective is to understand how the PAS score could guide a trustor's decision-making.

- Hypothesis 1: In a context of arbitrage between human agent and automaton agent, PAS is positively correlated with trust in automation.
- Hypothesis 2: In a context of arbitrage between human agent and automaton agent, PAS is not correlated with trust in human.
- Hypothesis 3: In a context of arbitrage between human agent and automaton agent, high PAS score predicts a preference to choose the automated system.

Participants

This study includes 40 participants (38 males and 4 females) between 19 and 28 years of age ($M = 23.3$, $SD = 2.25$). Participants are students of the École de l'Air et de l'Espace (French Air Force Academy) in Salon-de-Provence. All participants reported having normal or corrected vision and no history of neurological disorders and were naive to the study's hypotheses.

Tasks

Before starting the experiment, each participant completes a self-assessment scale called PAS. During the experiment, participants are placed in the context of a CAS (Close Air Support) mission, which is an air action performed by aircraft against hostile targets on the ground and in the vicinity of friendly forces. The objective of the CAS mission is the detection of an enemy target on the ground. In this experiment, the participant receives pictures of possible targets from two agents: a human agent ("JTAC") and an automation ("MAPS"). The JTAC (Joint Terminal Attack Controller) is a human agent that guides the aircraft from the ground to assist the pilot in detecting the enemy target. The MAPS system is similar to the Google Earth software, and is able to visualize the Earth using aerial images. The participant is responsible for selecting one of the two pictures that is for her/him the more probable target.

During each trial, the target description and the reliability of the two agents are first presented to the participant. Then two satellite images provided, one by the JTAC and the other one by the MAPS agent are displayed. The two satellite images are conflicting because they do not display the same location on the ground while they both match the description of the target (for example two different bridges). The participant's main task is to click on the satellite image that he or she believes corresponds to an enemy target. Once the choice is made, the participant is asked to answer three confidence scales regarding their choice. The level of confidence in each agent and the level of confidence in the answer is asked. Once the volunteer has answered all three statements, they can move on to the next trial. The task contains 64 trials per participant.

Measures

The experiment takes place on a 15-inch computer. The reliability and the nature of the agent are the independent variables. Four reliability values are proposed including 20%, 50%, 70% and 90%. Two types of agents are proposed, including a human agent (JTAC) or an automation (MAPS). Sixteen combinations are distributed as 4 (human reliability) x 4 (automation reliability). Four trials per combination, i.e. 64 trials per participant, are tested during the experiment. A database of 64 pairs of images is used for the experiment. The database comes from a previously conducted experiment that ensures both images equally refer to the proposed description. The experiment lasts on average 27 minutes. The dependent variables are trust in the human, trust in the automation, trust in the decision, response time and the participant's decision.

Perfect Automation Schema

The PAS scale consists of four items about high performance expectations and three items about all-or-nothing beliefs (Merritt et al., 2015). Each item is measured with a five-point Likert scale ranging from 'strongly disagree' to 'strongly agree'. The PAS scale appears to play an important role in understanding how trust works. Therefore, in this research, the impact of the PAS on the understanding of trust mechanisms is investigated.

Trust

The scale used is based on the Merritt confidence scale (Merritt et al., 2013). The participant responds to a statement such as "I trust JTAC" (trust measure 1) or "I trust MAPS" (trust measure 2) or "I trust my answer" (trust measure 3). They are provided with a non-scaled scale ranging from strongly disagree to strongly agree to respond.

RESULTS

The first result concerns the possible relationship between the PAS score and the trust granted to each agent, as reported by the trust measures 1 and 2. A Pearson correlation matrix is performed between the score PAS, the trust in the automation and the trust in the human. According to Table 1, the PAS score is positively correlated with trust in automation, $r(38) = 0.117$, $p =< 0.001$. This result supports hypothesis 1. However, no correlation between PAS and trust in humans is observed, $r(38) = 0.028$, $p = 0.163$. This result is consistent with hypothesis 2.

Then, the results concern the possible correlation between the PAS score and choices made by the participant. An Analysis of Variance (ANOVA) is performed on the participant decision variable with the dependent variable of PAS. This analysis reveals a significant effect between the two variables,

Table 1. Correlation matrix of PAS, trust in automation and trust in human.

		PAS	Trust in automation	Trust in human
PAS	r de Pearson	—		
	valeur p	—		
Trust in automation	r de Pearson	0.117 ***	—	
	valeur p	<.001	—	
Trust in human	r de Pearson	0.028	0.038	—
	valeur p	0.163	0.052	—

Note. *p <. 05, **p < .01, ***p < .001

$F(1, 2551) = 15.9, MSE = 109.85, p =< 0.001$). A Tukey post-hoc test is performed. The post test is significant ($p =< 0.001$): HumanChoice ($M = 16.7$, $SD = 0.0666$) vs AutomationChoice ($M = 17.1$, $SD = 0.0833$). The participant's decision is therefore related to the score PAS, while participants with higher PAS score more often choose the picture provided by the MAPS agent than participants with lower PAS score. Hypothesis 3 is accepted.

DISCUSSION AND PERSPECTIVES

The aim of this experimental step was therefore to test whether the personality of trustors had a similar influence on their trust in automation and their interpersonal trust in a situation where they had to make a decision according to their trust in their two partners. To this end, an arbitrage situation was designed in which the trustors' PAS score was compared to the perceived trust in the machine (Hypothesis 1) and the perceived trust in the human partner (Hypothesis 2). Then these elements were put into perspective with the arbitration carried out by the trustors between the information transmitted by a human partner and that transmitted by an automaton agent to make a decision (Hypothesis 3).

Results concerning hypothesis 1 confirm those obtained by the previous authors, i.e. that the PAS score is positively correlated to the TA, even in an arbitration context. This confirms the interest of the method to profile individuals according to their relationship to automation. Moreover, this score is predictive of the trustor's attitude in an arbitrage situation (hypothesis 3). On the other hand, the PAS score is not correlated with the perceived trust in the human agent (hypothesis 2).

These elements confirm that the PAS is relevant for the context for which it was designed, but that its use cannot be extended to another type of trust (here IT).

Hence, the next relevant step is to design a generic model of partner acceptability evaluation score compatible with all possible trust agents, i.e. a robot agent (TA), a human agent (IT), but why not also an artificial intelligence (trust in AI) and oneself (self-confidence). It is for this purpose that the extension of the *TorTeeX* model will be mobilized.

Taken together, the theoretical confrontation of the different conceptualizations of trust and the results obtained here allow us to propose a line of thought: trust would be a cognitive process that consists in evaluating the reliability of a resource in order to anticipate possible futures. In this sense, it would be a process that contributes to decision-making, in particular in the anticipation process.

ACKNOWLEDGMENT

The authors would like to acknowledge the French 'AID' (Agence de l'Innovation de Défense) and 'Dassault-Aviation' for their funding of the PhD defense program.

REFERENCES

Bollon, F. (2020). Recognized Cyber Picture et Confiance interpersonnelle : Optimiser les activités collaboratives cyber militaires. Thèse de doctorat, Université de Bordeaux. in French.

Bollon, F., Marchand, A. L., Maille, N., Blättler, C., Chaudron, L., & Salotti, J. M. (2019, October). Interpersonal trust to enhance cyber crisis management. In HFES-Europe 2019-Human Factors and Ergonomics Society, Europe Chapter.

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.

Chaudron, M. L., Chouchane L. & Burkhardt, J-M. (2022, June 1–2) A Generic Trust Model. International Conference on Cognitive Aircraft Systems ICCAS 2022. ISAE-SUPAERO (Ed.). Toulouse, France.

Chouchane, L. (2022). Trust Mobility Teaming. Master 2 thesis, UGE & AMU & École de l'air et de l'espace. in French.

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), 909–927.

De Jong, B. A., Dirks, K. T., & Gillespie, N. (2016). Trust and team performance: A meta-analysis of main effects, moderators, and covariates. Journal of applied psychology, 101(8), 1134.

Deutsch, M. (1973). The resolution of conflict: Constructive and destructive processes. Yale University Press.

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), 611–628.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. Human factors, 44(1), 79–94.

Endsley, M. R. (1995). Measurement of situation awareness in dynamic systems. Human factors, 37(1), 65–84.

Endsley, M. R. (2017). Toward a theory of situation awareness in dynamic systems. In Situational awareness (pp.9–42). Routledge.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. Human factors, 53(5), 517-527.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human factors, 57(3), 407–434.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: a failure to disagree. American psychologist, 64(6), 515–526.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. International journal of human-computer studies, 40(1), 153–184.

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

Lewandowsky, S., Mundy, M., & Tan, G. (2000). The dynamics of trust: comparing humans to automation. Journal of Experimental Psychology: Applied, 6(2), 104.

Lyons, J. B., & Guznov, S. Y. (2019). Individual differences in human–machine trust: A multi-study look at the perfect automation schema. Theoretical Issues in Ergonomics Science, 20(4), 440–458.

Lyons, J. B., Ho, N. T., Fergueson, W. E., Sadler, G. G., Cals, S. D., Richardson, C. E., & Wilkins, M. A. (2016). Trust of an automatic ground collision avoidance technology: A fighter pilot perspective. Military Psychology, 28(4), 271–277.

Lyons, J. B., Ho, N. T., Van Abel, A. L., Hoffmann, L. C., Sadler, G. G., Fergueson, W. E., ... & Wilkins, M. (2017). Comparing trust in auto-GCAS between experienced and novice air force pilots. ergonomics in design, 25(4), 4–9.

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: an integrative review. Theoretical Issues in Ergonomics Science, 8(4), 277–301.

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

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. Human factors, 50(2), 194–210.

Merritt, S. M., Unnerstall, J. L., Lee, D., & Huber, K. (2015). Measuring individual differences in the perfect automation schema. Human factors, 57(5), 740–753.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. International journal of man-machine studies, 27(5-6), 527-539.

Muir, B. M., & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. Ergonomics, 39(3), 429–460.

Parasuraman, R., E. Visser, M. Ming-Kuan Lin, and P. M. Greenwood. 2012. "Dopamine Beta Hydroxylase Genotype Identifies Individuals Less Susceptible to Bias in Computer-Assisted Decision Making." PLoS One 7: e39675, 1–9.

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. Journal of personality and social psychology, 49(1), 95-112.

Riley, V. A. (1994). Human use of automation. University of Minnesota.

Rotter, J. B. (1980). Interpersonal trust, trustworthiness, and gullibility. American psychologist, 35(1), 1.

Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task. Human factors, 49(1), 76–87.

Schoorman, F. D., Mayer, R. C., & Davis, J. H. (2007). An integrative model of organizational trust: Past, present, and future. Academy of Management review, 32(2), 344–354.