Proceedings of the Human Factors and
Ergonomics Society Annual Meeting
2023, Vol. 67(1) 139–144
Copyright © 2023 Human Factors
and Ergonomics Society
DOI: 10.1177/21695067231192294
journals.sagepub.com/home/pro
Sage

Explaining Trust Divergence: Bifurcations in a Dynamic System

Mengyao Li[1], Sofia I. Noejovich[1], Ernest V. Cross[2], and John D. Lee[1]

Abstract

When people experience the same automation, their trust in automation can diverge. Prior research has used individual differences—trust propensity and complacency—to explain this divergence. We argue that bifurcation as an outcome of a dynamic system better explains trust divergence. Linear mixed-effect models were used to identify features to predict trust (i.e., individual differences, automation reliability, and exposure). Individual differences associated with trust propensity and complacency increases the R^2 of the baseline model by 0.01, from $R^2 = 0.40$ to 0.41. Furthermore, the Best Linear Unbiased Predictors (BLUPS) for random effect of participants were uncorrelated with trust propensity and complacency. In contrast, modeling trust divergence from a dynamic perspective, which considers the interaction between reliability and exposure along with the individual by-reliability variability fit the data well ($R^2 = 0.84$). These results suggest dynamic interaction with automation produce trust divergence and design should focus on state dependence and responsivity.

Keywords

Trust in automation, Dynamic system, Individual difference, Trust dynamics

Introduction

As intelligent agents become increasingly autonomous on progressively more complex tasks, trust becomes more essential to designing effective human-automation cooperation (Chiou & Lee, 2021). Trust, defined as "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), is crucial for ensuring appropriate reliance on automation and avoiding its misuse, disuse, or abuse (Parasuraman, 1997). Often, people's trust in automation often evolves and converges to a relatively homogeneous level of trust. However, trust can also diverge. Interacting with the same automation, some people might develop high levels of trust whereas others might grow to distrust it (Kamaraj et al., 2023; Liu et al., 2021). This divergent trust is an interesting form of trust miscalibration because it describes how some people might over-trust and others might under-trust the same automation.

It might be useful to consider trust divergence as qualitative changes or 'bifurcation' in how people experience and trust automation. Bifurcation, well-studied in dynamical systems, describes how a small initial change of a system made to the parameter values, known as bifurcation parameters, can cause a sudden topological change in its behavior. In the context of trust in automation, this small initial change is often considered as differences in initial trust and individual differences, as well as variance in their initial perception and interaction with the automation. The bifurcation parameter refers to the changes in automation characteristics, such as an error. Previous research has highlighted various reactions following automation failures, including disbelievers, Bayesian decision-makers, and oscillators (Bhat et al., 2022). While researchers often rely on individual differences to explain diverse group behaviors. Yet, focusing solely on individual behaviors neglects the temporal aspect of how initial individual differences compound with the subsequent experiences of automation characteristics, especially when encountering the 'bifurcation parameter' (e.g., automation errors). The underlying mechanism contributing to the stabilized and diverging trust has received little attention and merits investigation. Three factors, namely individual differences, automation characteristics, and trust dynamics, may account for the trust bifurcation. In this paper, we argue that adopting the concept of bifurcation as an outcome of a

[1] Department of Industrial and Systems Engineering, University of Wisconsin – Madison, WI, USA
[2] Charles River Analytics, Cambridge, MA, USA

Corresponding Author:
Mengyao Li, Department of Industrial and Systems Engineering, University of Wisconsin – Madison, WI, USA.
Email: mengyaoli9687@gmail.com

dynamic system offers a more suitable framework for explaining trust divergence.

Individual Differences

The wide range of individual differences, encompassing backgrounds, personalities, and knowledge of automation, contributes to the variability in individuals' propensity to trust automation. Those with a higher inclination to trust may experience a greater decline in trust when interacting with low-performing automation (Merritt & Ilgen, 2008). Moreover, individuals who with a stronger "perfect automation schema" demonstrated greater declines in trust when they encountered automation errors (Dzindolet et al., 2002). Additionally, individuals also vary in automation-induced complacency, which can manifest as either a failure to detect or an delayed response to detecting errors (Bailey & Scerbo, 2007; Merritt et al., 2019). Prior research has found that complacency interacts with automation characteristics: the higher the system reliability, the more likely the operators become complacent (Parasuraman et al., 1993). Minor differences in individuals can influence the initial level of trust and subsequently shape the interpretation of new information. Thus, individual differences can influence trust divergence.

Hypothesis 1: Individual differences predict diverging of trust in automation.

Automation Reliability and Exposure

Because trust calibration is the correspondence between a person's trust in automation and the automation's capabilities, it has been consistently shown that automation capability significantly influences trust in automation (Dzindolet et al., 2002). Automation failures often have a much stronger influence on trust than automation successes: trust is difficult to build but can be lost quickly (Dzindolet et al., 2003; Manzey et al., 2012). Trust is continuous process influenced by the trust of a previous moment (Yang et al., 2023). Exposure to automation reflects the extent to which individuals have encountered and interacted with automated systems. Repeated exposures can have both positive and negative effects on individuals' behaviors and trust in automation. On one hand, repeated exposure can increase familiarity, indirectly influencing trust (Mayer et al., 1995). On the other hand, repeated exposures, especially with highly reliable automation, can induce complacency and decreased situational awareness, resulting in over-reliance on automation and over-react to automation errors (Dzindolet et al., 2002). Thus, the automation capability and exposure to automation can be potential causes of the diverging levels of trust and motivate the second hypothesis.

Hypothesis 2: Automation reliability and exposure predict diverging of trust in automation.

Trust Dynamics

Trust is inherently dynamic. People calibrate their trust over time as a continuous cognitive process (Gao & Lee, 2006). While researchers have highlighted the continuous and temporal elements of trust dynamics (Yang et al., 2023), limited past research has used trust dynamics to explain people's divergent opinions on automation. Using trust dynamics, trust divergence can be modeled as a bifurcation in a dynamic system: a small change in the initial state gradually influences behavioral framing and subsequent decision-making processes. This bifurcation results in trust stabilizing as two distinct trajectories. For example, in supervisory control, the individual differences shape the decision between manual control and automation. Once either decision is selected, it would provide positive or negative experiences. The experiences create inertia to keep people only focusing on either the advantages or disadvantages. Automation failures can be bifurcation transient point, which leads to trust divergence and long-term maintenance in certain states. Thus, the structural changes of the bifurcation depend on the combination of individual differences, the automation performance and the exposure, and their interaction over time, rather than on any individual factor alone.

Hypothesis 3: Trust is a dynamic system. People's varying responses to the interaction of automation characteristics and exposure predict diverging of trust in automation.

Method

The study was a 2 (reliability) × 2 (cycles) × 3 (events) within-subject study. Participants performed 12 decision-making tasks associated with managing a system of a simulated space station: the Habitat's Carbon Dioxide Removal System (CDRS). Participants were assisted by a conversational agent (Bucky) with 2 levels of reliability (i.e., high, and low). Each level of reliability had 2 repeated cycles of the CDRS tasks, each including 3 events (i.e., startup, venting, shutdown). Details of the study were documented in (Li et al., 2022).

Participants

A total of 24 participants (18 female, 6 male) were recruited (M = 23.7, SD = 3.6). Recruitment inclusion criteria included that participants should be comfortable using a computer and a touch screen interface as well as have some technical background (e.g., completion of engineering or science courses). Due to the safety concerns of COVID-19, the study took place online. It was a two-session, two-day study with each session lasting up to two hours. In total, the study lasted approximately four hours for each participant. Participants received $30 per hour for up to $120 for four hours of participation.

Procedures

After signing the consent form, participants completed a two-part training: the first provided a study overview and training on the CDRS system, while the second included an interactive demonstration of working with Bucky on decision-making in PRIDE. During the study, participants had 25 minutes to use the CDRS system to remove CO_2 from Habitat's environment by running the CDRS through three events (startup, venting, and shutdown) before their crew experienced CO_2 poisoning. For each event, the participant made two essential decisions with Bucky's aid. The first was selecting a procedure to run to remove the CO_2. Bucky recommended a procedure. The participant could either accept Bucky's recommendation or reject it and choose a different procedure. The second decision was deciding whether to rerun the procedure selected. As part of this decision participants would be advised by Bucky if the state of the CDRS was incorrect and if a different procedure should be run. The participants could either accept Bucky's recommendation or reject Bucky's recommendation and run a different procedure. The participants made their decisions either based on their knowledge from their training session or by relying on Bucky's recommendation. Once the procedure was selected, PRIDE automated the procedure execution. If the participant selected the incorrect procedure, an error occurred. The participant then had to manually stop the procedure and reselect a procedure. The participant finished the event by confirming the procedure ran correctly and completed the trust ratings.

Data Analysis

Linear mixed effect models identified features predicting trust as measured by the 12-item, 7-point Likert scale (Jian et al., 2000). To test our hypotheses regarding how individual differences, automation characteristics, and dynamics explain trust, we gather relevant features for each hypothesis.

For individual differences, we measured people's automation complacency and propensity to trust. We adopted the Automation-Induced Complacency Potential-Revised scale (AICP-R) (Merritt et al., 2019), which is a 10-item with response options on a five-point Likert scale ranging from 1 (strongly agree) to 5 (strongly disagree). Example items include, "Constantly monitoring an automation is a waste of time." For propensity to trust, we measured people's general tendency to trust automation using the Propensity to Trust Machines questionnaire (Merritt, 2011). This scale consists of six items with response options ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). Example items include, "I usually trust machines until there is a reason not to."

Automation characteristics were modeled as reliability condition and exposure. Reliability is a binary indicator of agent performance. Exposure is defined as the number of times participants experience the same automation characteristics, which is the number of cycles participants experienced.

Table 1. Mean Trust Values between Reliability and Exposure.

Reliability	Exposure	M(*SD*)	CI
High	1	5.52 (.19)	[5.14, 5.90]
	2	5.77 (.18)	[5.39, 6.14]
Low	1	4.42 (.25)	[3.90, 4.94]
	2	3.86 (.27)	[3.31, 4.41]

For the trust dynamics hypothesis, we considered the interaction of automation characteristic and exposure along with individuals' varying responses to the experiences.

Results

The mean trust score for the high-reliability condition was 5.78 (SD = 0.86) whereas, for the low condition, the mean trust score was 4.37 (SD = 1.44). The difference in mean trust values across reliability and exposure were shown in Table 1. From Figure 1 we observed that the path taken by individuals throughout the experiment was highly variable: some maintained a steady level of trust throughout the experiment, while others had dramatic drops in trust. The black lines represent six participants: three with the highest standard deviation and three with the lowest standard deviation in mean trust. The difference in paths reveals a divergence in trust when participants experience the low-reliability condition.

In Table 2, four linear mixed-effects models were built. Models were evaluated using the root mean square error (RMSE), Akaike information criterion (AIC), Bayesian information criterion (BIC), and conditional R^2 value. RMSE reflects the difference between predicted and actual values. AIC and BIC reflect how well the model fits the data with a term that penalizes model complexity. The lower these three metrics, the better the model performance. The conditional R^2 is the proportion of total variance explained by the model. The higher the R^2, the better the model performance.

Model 0 (Reliability | ID) uses ID as the random intercept serving as a baseline model which accounts for the overall trust level due to general individual differences.

Model 1 corresponds to the first hypothesis and tests the effects of specific individual differences on trust in automation, which were measured using automation-induced complacency and propensity to trust scales. The individual measures only slightly improved the marginal R^2 value. The effect of complacency and propensity are both statistically non-significant ($p = 0.61, p = 0.38$).

Model 2 corresponds to the second hypothesis and tests the effects of automation characteristics on trust. We used automation reliability and the number of cycles as exposure to automation. We added reliability and exposure as fixed effects to determine if the model performance would be improved. The effect of reliability [low] is statistically significant and negative, β = -1.40, 95% CI [-1.62, -1.19], $t(210) = -12.82$, $p <.001$; Std. β = -1.07, 95% CI [-1.38, -0.75], whereas the effect of exposure is non-significant, $t(210) = -0.69$, $p = 0.49$.

Model 3 corresponds to the third hypothesis and tests the effect of trust dynamics by adding the interaction between reliability and exposure along with the individual by-reliability variability. By adding the individual by-reliability variability, model 3 shows diverging effects in Figure 1 and would serve as the baseline model. The total explanatory power of this model is substantial with a high conditional R^2 value (0.84) and the part related to the fixed effects alone. Additionally, the AIC and BIC are the lowest for model 4, which indicates that the trust dynamic model explains the greatest amount of trust variation using the fewest possible parameters. Within this model, the effect of low reliability is statistically significant and negative, β = -1.10, 95% CI [-1.51, -0.69], $t(207) = -5.24$, $p <.001$; Std. β = -0.84, 95% CI [-1.16, -0.53].

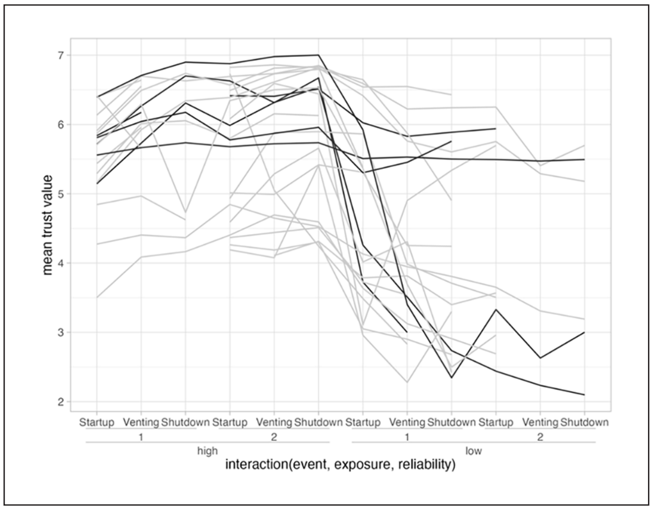

Figure 1. Trust diverges when people experience low-reliability automation.

The effect of exposure is statistically significant and positive, β = 0.24, 95% CI [0.05, 0.44], $t(207) = 2.43$, $p =.02$; Std. β = 0.19, 95% CI [0.04 0.34]. The interaction effect of the exposure and reliability is statistically significant and negative, β = -0.80, 95% CI [-1.14, -0.47], $t(207) = -4.70$, $p <.001$; Std. β = -0.62, 95% CI [-0.88, -0.36]. Trust in the high-reliability condition is an estimated 4.89 on a Likert scale of 7. The trust score is 1.10 points lower in the low condition, 0.24 points higher in the second exposure, and 0.81 points lower if there is an interaction between the low condition with the second exposure. For the random effects, the standard deviation for by-subject random intercepts indicates that trust levels for subjects varied around the average intercept of 0.69 points by about 0.77 points. Additionally, we used Best Linear Unbiased Predictions (BLUPs) to predict random effects and found no correlations with the automation complacency ($R^2 < 0.01$) and propensity to trust ($R^2 = 0.03$). These results again validate that individual differences do not account for trust divergence and supports the trust dynamics hypothesis.

Discussion

We observed that trust diverges when people experienced automation error: some people maintained a steady level of trust whereas others showed a drastic decline in trust. To explain this trust divergence, we evaluated three hypotheses—individual differences, automation characteristics, and trust dynamics–using linear mixed effects models. We found that the trust dynamics model, which uses automation exposure and reliability as an interaction fixed effect, with individual differences and participants as a random intercept and slope, yielded the highest R^2 and lowest AIC and BIC values. Results suggest that the trust dynamics model best explained the trust divergence. Because trust dynamics consider individual differences and how people's trust is reinforced by the automation characteristics and multiple exposures over time. Our results reinforce the notion that individual differences alone are insufficient to explain trust divergence. Instead, the concept of bifurcation in a dynamic system may provide a better explanation. This concept

Table 2. Performance metrics comparison between regression models.

#	Model	Formula	RMSE	AIC	BIC	R^2 (cond.)
0	Baseline model	$trust \sim 1 \mid ID$	0.97	671.38	681.49	0.40
1	Individual differences	$trust \sim complacency + propensity + (1 \mid ID)$	1.03	613.94	630.15	0.41
2	Automation reliability and exposure	$trust \sim reliability + exposure + (1 \mid ID)$	0.70	555.31	572.16	0.68
3	Trust Dynamics	$trust \sim reliability + exposure + reliability * exposure + (reliability \mid ID)$	0.48	481.13	508.09	0.84

describes how even slight changes in a system can lead to qualitatively different behavior, which might correspond to certain individuals maintaining stable trust in automation while others experience sudden shifts in trust.

Whether trust diverge reflects enduring traits or states that emerge from automation interaction has major system design implications. These mechanisms parallel those associated with the concept of "accident proneness." Prior studies found that individuals who have experienced incidents of accidents in the past are more likely to experience them in the future than are individuals who have not experienced an accident (Bates & Neyman, 1952). Heckman argued that this conditional probability of accident proneness is based on structural relationships of state dependence, rather than heterogeneity in population and individual differences (Heckman, 1981). Enduring individual differences or traits suggest an emphasis on selection in system design, whereas state dependence would emphasize interaction design.

Interaction design from the trust dynamic perspective suggests that systems should measure and manage trust across human-automation interactions. Rather than focusing only on generic trust calibration through more transparent designs, a dynamic perspective suggests a focus on "responsitivity", where the automation detects and responds to changes in trust (Chiou & Lee, 2021). The importance of a dynamic perspective is even more important in hybrid teams with more than one human operator interacting with the automation. In these teams of over- and under-trust can circulate as a contagion within the network. Trust circulates through the network via explicit communication or implicit observations of others' interactions and norms (Stewart, 2003). Drawing inspiration from the widely used Susceptible-Infectious-Recovered (SIR) dynamic system model in epidemiology, researchers can explore the influence of network dynamics on trust bifurcation (Nakahara & Doya, 1998). Gorman and colleagues have previously conceptualized teams as dynamic systems, revealing the importance of concepts like attractors and synchronization (Gorman et al., 2017). Future research can understand and model trust dynamics in a hybrid team, identifying the roles and impacts of attractors, perturbation, and synchronization.

Our findings on trust dynamics conforms with the state dependence theory (Heckman, 1981). When designing the system, it is crucial adopt a state-dependent and dynamic perspective to evaluate human performances and trust. Early-stage measurement of trust and identification of distinct populations experiencing divergent trust patterns can inform the development of personalized systems to manage trust more effectively.

Conclusion

Even when people experience the same automation, their trust in automation can diverge over time. Prior research has typically focused on individual differences to explain trust divergence. However, we showed that trust divergence was best modeled by trust dynamic perspective, which considers the interaction between reliability and exposure along with the individual by reliability variability ($R^2 = 0.84$). Our results suggest the concept of bifurcation in dynamic systems, which describes how small changes in a system lead to sudden shifts in behavior, might explain trust divergence.

Acknowledgments

This work was supported by NASA Human Research Program No.80NSSC19K0654. We would like to thank members of the University of Wisconsin-Madison Cognitive Systems Laboratory for their insightful discussions and comments.

ORCID iD

Mengyao Li https://orcid.org/0000-0002-0819-4693

References

Bailey, N. R., & Scerbo, M. W. (2007). Automation-Induced Complacency for Monitoring Highly Reliable Systems: The Role of Task Complexity, System Experience, and Operator Trust. *Theoretical Issues in Ergonomics Science*.

Bates, G. E., & Neyman, J. (1952). Contributions to the Theory of Accident Proneness. *University of California Press*.

Bhat, S., Lyons, J. B., Shi, C., & Yang, X. J. (2022). Clustering Trust Dynamics in a Human-Robot Sequential Decision-Making Task. *IEEE Robotics and Automation Letters*, *7*(4), 8815–8822.

Chiou, E. K., & Lee, J. D. (2021). Trusting Automation: Designing for Responsivity and Resilience. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *65*(1), 137–165.

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.

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: The Journal of the Human Factors and Ergonomics Society*, *44*(1), 79–94. https://doi.org/10.1518/0018720024494856

Gao, J., & Lee, J. D. (2006). Extending the Decision Field Theory to Model Operators' Reliance on Automation in Supervisory Control Situations. *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans*, *36*(5), 943–959.

Gorman, J. C., Dunbar, T. A., Grimm, D., & Gipson, C. L. (2017). Understanding and Modeling Teams as Dynamical Systems. *Frontiers in Psychology*, *8*.

Heckman, J. J. (1981). Heterogeneity and State Dependence. In *Studies in Labor Markets* (pp. 91–140). University of Chicago Press.

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71.

Kamaraj, A. V., Lee, J., Parker, J., & Domeyer, J. E. (2023). Bimodal Trust: Relationship Between Drivers' Trust in Reliable Automation and Response to a Surprise Automation

Error. *In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 67*.

Lee, J., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, *46*(1), 50–80.

Li, M., Erickson, I., Cross, E., & Lee, J. (2022, October 17). Estimating trust in conversational agent with lexical and acoustic features. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*.

Liu, J., Akash, K., Misu, T., & Wu, X. (2021). Clustering human trust dynamics for customized real-time prediction. *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, 1705–1712.

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human Performance Consequences of Automated Decision Aids: The Impact of Degree of Automation and System Experience. *Journal of Cognitive Engineering and Decision Making*, *6*(1), 57–87.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. In *Source: The Academy of Management Review* (*Vol. 20*, Issue 3, pp. 709–734).

Merritt, S. M. (2011). Affective Processes in Human–Automation Interactions. *Human Factors*, *53*(4), 356–370.

Merritt, S. M., Ako-Brew, A., Bryant, W. J., Staley, A., McKenna, M., Leone, A., & Shirase, L. (2019). Automation-Induced Complacency Potential: Development and Validation of a New Scale. *Frontiers in Psychology*, *10*, 225.

Merritt, S. M., & Ilgen, D. R. (2008). Not All Trust Is Created Equal: Dispositional and History-Based Trust in Human-Automation Interactions. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(2), 194–210.

Nakahara, H., & Doya, K. (1998). Near-Saddle-Node Bifurcation Behavior as Dynamics in Working Memory for Goal-Directed Behavior. *Neural Computation*, *10*(1), 113–132.

Parasuraman, R. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors*, *39*(2), 230–253.

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance Consequences of Automation-Induced "Complacency." *The International Journal of Aviation Psychology*.

Stewart, K. J. (2003). Trust Transfer on the World Wide Web. *Organization Science*, *14*(1), 5–17.

Yang, X. J., Guo, Y., & Schemanske, C. (2023). From Trust to Trust Dynamics: Combining Empirical and Computational Approaches to Model and Predict Trust Dynamics in Human-Autonomy Interaction. In V. G. Duffy, S. J. Landry, J. D. Lee, & N. Stanton (Eds.), *Human-Automation Interaction: Transportation* (pp. 253–265). Springer International Publishing.