A two-factor model of trust and distrust for new technology environments.

Lisa Z. Wise, Samuel J. McKay and Jason L, Skues, Department of Psychological Sciences, Faculty of Health Arts and Design, Swinburne University of Technology, VIC, Australia.

1. Trust in new technologies

The concept of trust is important in the context of new technologies designed to incorporate automated and autonomous capabilities into defence operations. In situations where operator trust in automation is not appropriately calibrated, over and under reliance on such systems can lead to potentially drastic negative consequences such as missing important changes in the environment or the disuse of useful systems [1]. This paper proposes a two-factor model of trust and distrust within a cognitive ergonomics framework. Within this framework, trust operates to reduce cognitive workload by delegating responsibility to a trustee, and distrust, as a non-pejorative construct, operates to increase metacognitive awareness through monitoring and validating the performance of the trustee. It should be noted that the monitoring and validating processes of distrust increases cognitive workload overall, but allows the opportunity to take over a task that has been previously delegated.

New automated and autonomous technologies present two significant challenges within the defence context:

1. ensuring that personnel can be trained with the appropriate trust-based cognitive skills to interoperate with automated and autonomous technology in their future operational roles; and
2. ensuring that the design of future systems is based on a clear theoretical understanding of the cognitive and metacognitive outcomes of promoting trust and distrust so as to support appropriate reliance on such systems.

Most current models of trust, while acknowledging the multidimensionality of trust, view trust and distrust as polar opposites within the same fundamental "trust/distrust" construct. Furthermore the literature does not clearly differentiate between the notions of dispositional trust, trust as an operator (allocation of trust by a trustor to a trustee to decrease complexity in the situational context), trust-based behaviour (behavioural implications of the current state of trust for both the trustor or trustee), nor does it adequately account for the context-dependence and dynamic nature of each of these factors. Furthermore, as noted by Abbass et al. [2], it may be possible to distinguish between trust between human and machine, trust between human and automation, trust between human and computer, and trust between human and robot.

This paper begins with a brief overview the concepts of trust and distrust in the management, social psychology and automation literature to refine a two-factor cognitive model of trust and distrust in complex and dynamic environments. It then proposes a program of research to allow for quantitative and qualitative tests of predictions of the two-factor model of trust in terms of cognitive behaviour such as shedding workload via automation, monitoring of on-going performance, choice of information sources etc.

2. Trust as a construct

Trust is an important construct in the social sciences because it increases human performance by reducing uncertainty and increasing efficiency, which leads to better quality interactions and improved task functioning [3]. In organisations, trust enhances intra- and inter-business relationships, supporting cooperation, delegation, social order, and control [1-5]. Within the automation space, trust allows for appropriate calibration of expectations around the capacities, uses and requirements of different

technologies, which leads to appropriate use and reliance on automation [6-7]. Despite the increasing importance of the concept of trust within the technology space, there is still no clear consensus on its definition, although across most theoretical orientations, trust is envisaged as a complex multidimensional construct which includes perceived vulnerability to risks related to the motivations, intentions and prospective actions of other agents. The notion of distrust is also seen to be complex and multidimensional, and may be influenced by different parameters other than trust [1-2, 8-9]. Few researchers propose two distinct factors of trust and distrust [1-2, 8], but instead view trust and distrust as comprising different facets of multidimensional trust, with trust attitudes and behaviours having a positive valence (confidence, faith, acceptance) and distrust attitudes and behaviours having a negative valence (suspicion, surveillance, rejection).

The model of trust and distrust proposed in this paper views trust and distrust as distinct and separate attitudinal constructs which give rise to different cognitive and meta-cognitive strategies and behaviours. The novel approach of this model is that the link between trust/distrust attitudes and trust/distrust behaviours can be either adaptive or maladaptive in given scenarios for both trust and distrust. This contrasts with most prevailing theoretical perspectives that view trust as exclusively positively-valenced and distrust as exclusively negatively-valenced. The primary motivation for this model of trust is to explore the role of trust in the delegation of responsibility and the distinct role of distrust in the process of monitoring and verification of performance such that trust acts as an operator to reduce cognitive workload, and distrust acts as an operator to maintain metacognitive awareness of current priorities in complex and dynamic contexts. For example, trust can function to reduce complexity in a situation in which an operator assumes automation is reliable and hands off tasks to the automated system. However, such a shedding of workload is beneficial only if the automation performs as expected, otherwise it may mean that the operator is overwhelmed by the cognitive workload required to identify and recover from performance errors made by automated systems.

While there may be a cognitive cost of undertaking the "distrust behaviours" of monitoring and verification of performance, distrust serves to reduce the perceived complexity of a situation by engaging operators with automated systems through the monitoring and verification of automation performance, ensuring that automation errors or issues can be anticipated. While operator engagement with a system involves a higher cognitive workload baseline than being complacent, if operators are actively engaged with monitoring and verifying performance, they are able to detect errors or issues earlier and thus have more opportunity for adaptation and mitigation of risks resulting from such errors [10]. In other words, having a higher baseline cognitive workload through active engagement with monitoring an automated system allows for risk mitigation to occur with lower cognitive cost and a reduced chance of being overwhelmed by a sudden increase in workload [10]. If on the other hand, the operator is overly suspicious of an automation system, even when it is highly reliable, then such behaviour will increase the workload of the operator, as they check for errors that are not present.

This model of trust and distrust is presented in Figure 1, showing trust and distrust as separate orthogonal constructs, with the combination of high trust and high distrust promoting the combination of delegation and verification of system performance in complex environments. Such a model has been developed in an interpersonal trust framework previously [8], but has not been used in the human factors and automation domains. Additionally, the inclusion of the new complexity and uncertainty aspects of the model allows for more nuanced and complete assessments of automation reliance and calibration of operator trust within differing environmental contexts and tasks. As tasks become more complex and uncertain, the ability to delegate responsibility (trust) or to take over a previously-delegated task in an appropriate manner becomes less and less possible without having monitored and verified ongoing performance of the delegated task (distrust). Distrust creates higher cognitive workload but allows for the possibility of recovery if trust is broken.

Each of four quadrants in the model provides the opportunity to make predictions about behaviour and usage patterns with automation systems that have differing levels of reliability and trust related information provided to operators. Additionally, by adding the previously unmeasured context-related situational complexity and uncertainty dimensions, the model can be used to assess how operator strategies change as complexity of tasks increase or decrease, while controlling for operator trust and distrust levels in the automation. Assessing such changes is vital for developing a robust model because operators and personnel will be exposed to a wide variety of contexts and situations in their roles that

may have major impacts on the way they interact with automation systems. For example, in a diagnostic aid task [11], error detection of false alarms (incorrect detection of signal in environment) and misses (automation fails to detect signal in environment) had differing detection rates under high and low workload tasks. Under high workload misses were detected more readily than false alarms, while under low workload both were detected at equal levels. Within the model proposed in Figure 1, as workload increases, more delegation of responsibility to trusted systems is required, and it is important to be able to prioritise what to delegate, what to monitor, and the prioritisation of delegation and monitoring amongst different tasks and processes. The degree to which trust and distrust behaviours can be trained as cognitive skills is an open question to be pursued.

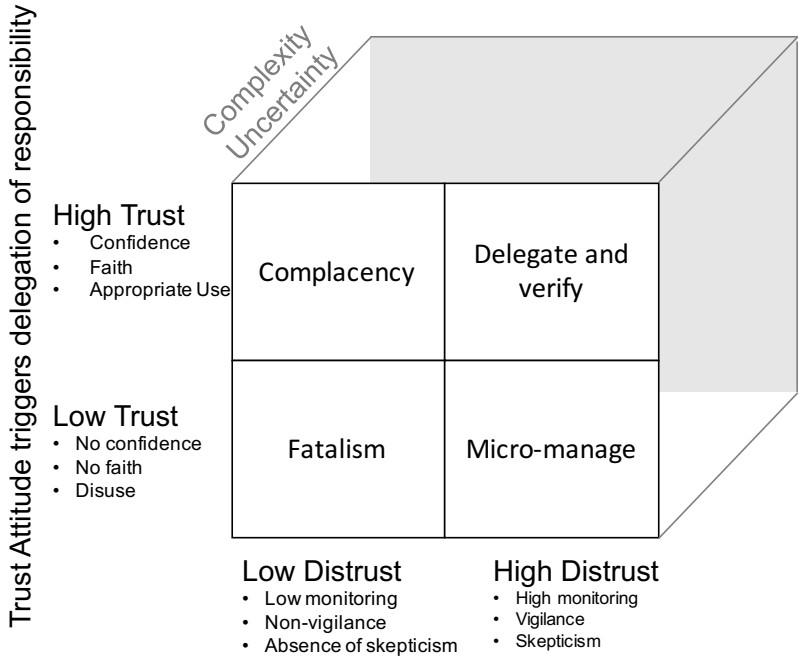

Figure 1. Left panel shows Trust and Distrust as two distinct factors leading to different trust behaviours: delegation and verification respectively.

3. Future Directions

The model of trust and distrust presented in this paper provides a qualitative framework for understanding trust/distrust in complex environments. In order to test predictions of the model of trust and distrust proposed, it is important for us to develop appropriate task environments and metrics to allow quantitative or computational analysis in addition to qualitative assessment of strategies and tactics. There multiple ways that such testing could be undertaken.

One option is to develop laboratory-based experimental tasks, an approach best suited to testing specific implicit and explicit aspects of trust/distrust attitudes and their effects on the choice of cognitive strategies and behaviours for specific task domains. Robust experimental paradigms for testing implicit attitudes, measuring response times and errors, and gathering data on dispositions, working memory capacity and mental workload can be readily programmed, for example using cognitive psychology research software such the Inquisit programming platform (http://millisecond.com). Another option is to select domain-specific commercial-off-the-shelf games that provide appropriate scenarios and levels of automation within simulated task domains and to develop methods of recording quantitative and qualitative performance data from the games to test hypotheses. The third option is to use an open-source research-oriented platform such as Aptima DDD

team-based simulation software (https://github.com/Aptima/DDD) to program appropriate scenarios and capture required experimental data. The advantage of this option is that the task domains are designed to simulate important aspects relevant real-world tasks and to capture data relating to team performance, and both the simulation software and the data capture processes are transparent and can be readily replicated by other researchers. Each of these options provides distinct benefits and it is likely that a combination of methods would be used to develop a complete experimental program.

A major implication of the proposed model of trust presented in this paper is that trust and distrust play a pivotal role in the moderation of cognitive workload in complex environments, and thus trust and distrust behaviours can be considered to be important and distinct cognitive skills that require specific training for complex future technology domains. Whilst being a trusting person may be construed as socially-desirable disposition in building a social community, appropriate levels of trust and distrust behaviours, as per Figure 1, are likely to play an important role in facilitating optimal performance in complex human-machine environments by moderating complexity, uncertainty and cognitive workload.

References

1. 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.
2. Abbass, H.A., Petraki, E., Merrick, K., HArvey, J., & Barlow, M. (2015). Trusted autonomy and cognitive cyber symbiosis: open challenges. *Cognition and Computation*, doi:10.1007/s12559-015-9365-5
3. Simpson, J.A. (2007). Foundations of interpersonal trust. In *Social Psychology: Handbook of Basic Principles (2nd edition*, pp 587-607.
4. Kramer, R.M. (1999). Trust and distrust in organisations: emerging perspectives, enduring questions. *Annual Review of Psychology, 50(1)*, 569-598.
5. Schoorman, F.D., Mayer, R.C., & Davis, J.H. (2007). An integrative model of organisational trust: past, present and future. *Academy of Management Review, 32(2)*, 344-354.
6. Lee, J.D., & See, K.A. (2004). Trust in automation: designing for appropriate reliance. *Human Factors, 57(3)*, 407-434.
7. Parasuraman, R., & Riley, V. (1997). Humans and automation: use, disuse, misuse and abuse. *Human Factors, 39(2)*, 230-253.
8. Lewicki, R.J., McAllister, D.J., & Bies, R.J. (1998). Trust and distrust: new relationships and realities. *Academy of Management Review, 23(3)*, 438-458.
9. Hoff, K.A. & Bashir, M. (2015). Trust in automation: integrating empirical evidence on factors that influence trust. *Human Factors, 57(3)*, 407-434
10. Young, M.S. & Stanton, N.A. (2002). Attention and automation: New perspectives on mental underload and performance, *Theoretical Issues in Ergonomics Science*, *3(2)*, 178-194, doi: 10.1080/14639220210123789
11. Dixon, S.R., Wickens, C.D., & Chang, D. (2005). Mission control of multiple unmanned aerial vehicles: a workload analysis. Human Factors, 47(3), 479-487. doi:10.1518/001872005774860005