Towards Trust-Aware Human-Automation Interaction: An Overview of the Potential of Computational Trust Models

Aya Hussein Sondoss Elsawah Hussein A. Abbass
School of Engineering and Information Technology
University of New South Wales, Canberra, Australia
a.hussein@student.adfa.edu.au, {s.elsawah, h.abbass}@adfa.edu.au

Abstract

Several computational models have been proposed to quantify trust and its relationship to other system variables. However, these models are still under-utilised in human-machine interaction settings due to the gap between modellers' intent to capture a phenomenon and the requirements for employing the models in a practical context. Our work amalgamates insights from the system modelling, trust, and human-autonomy teaming literature to address this gap. We explore the potential of computational trust models in the development of trust-aware systems by investigating three research questions: 1- At which stages of development can trust models be used by designers? 2- how can trust models contribute to trust-aware systems? 3- which factors should be incorporated within trust models to enhance models' effectiveness and usability? We conclude with future research directions.

1. Introduction

Recent technological advances have led to the inclusion of some levels of automation in a wide range of applications. The Industry 4.0 is seeing a shift in automation from a mere tool in the hand of human operators to accomplish a task, to machine autonomy, where automation is a cognitive artificial actor with the capacity to work with, or even replace, the human. Thanks to artificial intelligence, machines have become able to perform sophisticated high level tasks including planning and decision making. Nonetheless, fully autonomous machines that operate in dynamic and unstructured environments are not expected to be realised in the near future, due to their lack of human-like general intelligence. Aside from the technical challenges, it may not be desirable to give machines full autonomy due to ethical concerns and the need for accountability [1]. Thus, interaction schemes that combine the strengths of human and automation capabilities have been the focus of many recent studies.

Supervisory control is an interaction scheme that allows automation to operate at a higher level of autonomy while requiring the human to monitor its operation and intervene when needed. Hence, rather than acting as an operator, the human is asked to perform the role of a supervisor or a teammate who can take over control in unexpected events or intervene only when needed to modify goals. In such a setting, trust is vital for improving system performance due to its impact on human willingness to delegate tasks to automation. Besides, trust affects the frequency of human's intervention with automated tasks and his/her rate of acceptance of automation recommendations [2]. The reliance behaviour adopted by the human can eventually affect mission success and performance [3].

Both over-trust and distrust in automation can lead to catastrophic outcomes. For instance, human over-trust in automation was blamed for the death of the owner of a Tesla car in March 2018 [4]. This happened when the auto-pilot failed to recognise a concrete barrier, so the car veered off the highway, accelerated, and crashed into the barrier. Later, Tesla revealed that the driver had enough time to intervene to prevent the crash, but no action was taken. On the other hand, the disuse of reliable, though imperfect, automation results in dismissing its potential benefits. It has been warned that if the public rejected the auto-pilot, its safety levels would be dismissed causing the loss of about 900,000 lives that could otherwise be saved [5]. Thus, designing trust-aware human-automation interaction can be critical to mission success. Towards this end, many pieces of research have been devoted to studying factors that affect trust [6] as well as investigating methods for trust calibration [7, 8].

Several computational models have been proposed by past studies to quantify trust and represent its interaction with other factors. While different individual, cultural, environmental, and automation-related factors can affect trust and its role within a mission [2], existing trust models do not incorporate all these factors together, as such a

URI: https://hdl.handle.net/10125/63786
978-0-9981331-3-3

comprehensive model would require extensive human experiments to collect the data required for model calibration. Instead, many existing models capture the relationship between automation capabilities, trust, reliance rate, and system performance [9, 3, 10, 11]. In addition, some models include other variables like self-confidence [9], workload [12], and human expectations [3]. This raises a question about the truely important factors that need to be incorporated within a computational model for trust and that may affect the accuracy of its prediction.

Most of the computational models for trust have been validated by showing their ability to replicate data from human experiments [9, 13, 10]. The models were evaluated based on their accuracy of predicting subjects' trust ratings [13, 14], their rates of reliance on automation [9, 13, 10, 3, 11], and the overall performance [10, 3, 11]. Nevertheless, there are relatively few attempts to utilise these models towards enhanced human-automation interaction. This could be possibly due to the lack of clarity of the capabilities of computational trust models and the different ways in which these models can be utilised towards building trust-aware systems. That is, a designer may refrain from using trust models within the development of human-machine systems as he/she thinks these models are not practical and do not translate into specific actions or specific design decisions. Hence, it is important to characterise how computational trust models can be used as practical development tools rather than as a mere abstraction of the phenomenon. We argue that identifying the means of using such models by system designers and highlighting their potential in pushing forward trust-aware interactions would encourage the designers to utilise these models. Another possible reason why computational trust models are not widely employed by designers, could be related to the fact that most of the existing models ignore important factors (e.g individual skills and self-confidence) that can affect the dynamics of trust and reliance within the mission; which limits model usability.

Our work consider a performance-centeric view of trust where no deception or security breaches are expected from the machine. The objective of this work is twofold. First, this paper aims to characterise different categories of trust models and demonstrate how each category can be used within different stages of the development of trust-aware human-automation interaction. To our knowledge, this is the first attempt to investigate the possible ways in which a quantitative model for trust can be used within human-automation interaction settings. Second, this work aims to identify the key factors that should be included in computational trust models to boost their ability to closely reflect the interaction between system variables. Models including these factors are more likely to fulfill designers' needs in terms of having a holistic picture of the system.

2. Research Questions

In order to study the research objectives, we formulate three research questions which this paper contributes to their answer. The research questions are:

- At which stages of system development can computational trust models be used by system designers? To answer this question, section 4 distinguishes between two classes of models, offline and online models, and show the suitability of using these models within the design and deployment stages of system development.
- How do computational trust models contribute to the development of trust-aware systems? This question is further divided into two sub-questions:
  - How do trust models inform the *design* of trust-aware systems? To answer this question, section 5 uses insights from system modelling literature to show how computational trust models provide analytical tools that help designers identify design limitations and spot places for improvements.
  - How do trust models contribute to the *deployment* of trust-aware systems? The answer to this question is investigated in section 6 which brings diverse examples from the human-autonomy teaming literature to demonstrate the role of quantitative trust models in the deployment of adaptive systems. Figure 1 provides a summary of the research questions presented so far.
- What are the key factors that need to be incorporated within computational trust models to enhance their ability to meet the needs of system's designers? To study this question, section 7 examines the matching between the capabilities of trust models (in terms of how well they represent the system) and the system designers' needs (in terms of ensuring appropriate human reliance on automation and optimised system performance). Thus, we elicit the factors that need to be incorporated within trust models.

3. Key Concepts

Offline and online models: Existing trust models might

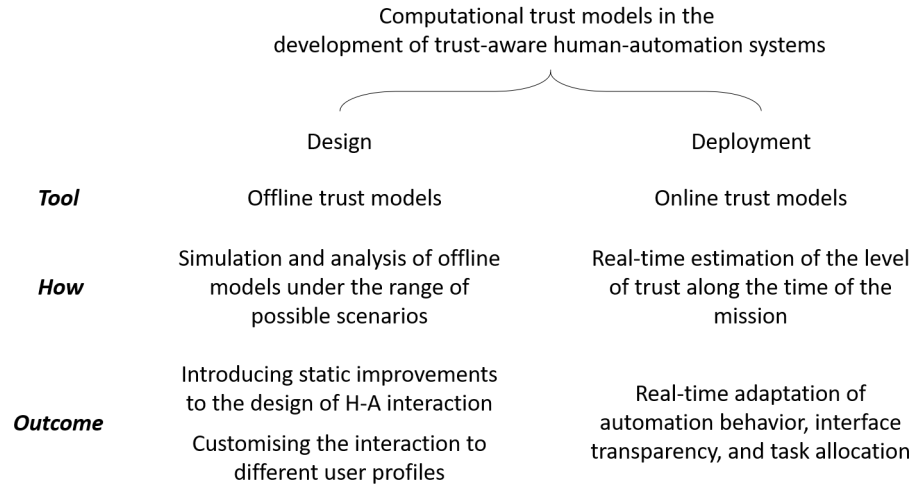

Figure 1. The role of computational trust models in the development of trust-aware human-automation interactions.

be classified based on different dimensions, for example probabilistic [13, 15, 16] versus deterministic [9, 12] or cognitive [3, 10] versus neural [17] models. However, to facilitate the investigation of our research questions, we propose classifying trust models into two classes based on the input data used to generate their predictions; these classes are offline and online models. *Offline models* use parameter values which are set a priori, i.e while the system is not under operation, to predict the behaviour of the modelled system. On the contrary, in addition to using some parameter values that are set a priori, *online models* make use of the observed data that is available during system operation to generate in-situ evidence-based predictions.

Customised improvements: we use this term to refer to design customaisation that is applied for a certain group of users based on their skills, abilities, or level of experience.

Adaptation: this term refers to the ability of the system to finely tune its default settings based on the actual scenario it is performing in. Three types of adaptations are studied in the human-autonomy literature: behaviour adaptation [18], transparency adaptation [16], and level of autonomy adaptation [19].

4. Offline and Online Models for Trust

As offline trust models are able to generate predictions, based on only an initial set of parameters, they are naturally suitable for being used within the design stage of system development. Offline models can be used to study system performance under different conditions and to gain deeper insights on how different factors interact to determine the human behaviour. Given a set of parameter values as input, an offline model of trust can be used to predict a full sequence of the levels of trust, rates of reliance on automation, and measures of performance over the time of the mission. Such models are usually based on feedback loops such that the state of the system, in terms of the values of its variables, at a given time step determines its state at the next time step [9, 3]. This can be useful to evaluate both performance trends and overall performance under different initial conditions. Therefore, offline models can be useful in evaluating alternative design options and to give an estimate of performance bounds.

A handful of offline models for human trust in automation has been proposed in previous studies. For instance, Gao and Lee [9] modelled the interaction between human and automation using a quantitative, feed-back based system. The model takes as input the initial levels of trust and self-confidence as well as the levels of automation and human capabilities. Trust and self confidence are modelled to change overtime according to the perceived performance of automation and human, respectively. The gap between trust and self confidence is then used to estimate the decision of reliance on automation. Lastly, the reliance decision determines whether the system performance in the next step will be determined by automation or manual performance.

Another example of offline models is presented in [3] in which the author proposed a system dynamics model for trust in human-automation interaction in scheduling tasks of collaborative multiple unmanned vehicles. In this model, the level of trust changes based

on the gap between the expected and the perceived performance. The rate of human interventions is modelled to be negatively related to the current level of trust. The negative effects of increased human workload on performance are captured by relating the human added value to the number of interventions using an inverted U-shaped table function. Both the automation capability and the human added value are used to estimate area coverage rate, which is in turn used to calculate the perceived performance.

On the other hand, online trust models rely on data extracted from the actual interaction to generate their prediction. Thus, they can be used to provide real-time estimation of the level of trust. In fact, most of the existing computational models for trust fall into the class of online models. Although trust can be measured directly through questionnaires, this can be impractical as it leads to frequent interruptions of task execution. Therefore using models for estimating trust is a more convenient alternative as it provides a non-intrusive way of estimating trust. Hence, online models are beneficial to sense changes in trust and respond accordingly.

Xu and Dudek [13] proposed a probabilistic model for trust based on dynamic Bayesian networks to predict the level of trust based on real-time data. The model uses both causal and evidential variables to infer the current level of trust. The causal variables used are the previous and current levels of automation performance and the previous level of trust. Meanwhile, the evidential variables are the current rate of interventions besides trust evaluation, if any. Xu and Dudek used subjects' data to train and evaluate their models such that a separate personalised model was used for each subject.

Nam et al. [15] proposed a probabilistic model for the real-time estimation of human trust in simulated robot swarms in target foraging missions. The model is based on the hypothesis that human trust in the swarm at a certain time step is a function of the swarm physical characteristics at this time step as well as the previous level of trust. To calibrate model parameters, Nam et al. used subjects' ratings of their level of trust along the experiment. General as well as personalised models have then been built to be used to predict trust given in-situ observed data about swarm parameters.

Offline and online models for trust use input data that can be available at the design and deployment stages of system development, respectively. This, in fact, highlights the opportunity of using these models in these two stages within the development of trust-aware systems. While insights obtained at the design stage can inform the design decisions and guide to ways of system improvements, the information obtained within the deployment stage can be used to adjust the system and finely tune its settings to suit the current situations.

5. Offline Models in the Design of Trust-Aware Systems

Offline models can be very useful in evaluating candidate design options by predicting their effects on performance under different possible contexts in which the system is intended to operate. In the conceptual model of trust in [2], Lee and See show that different individual, cultural, and environmental contexts can influence the formation of trust and the decision to rely on automation. Individual factors including age, propensity to trust, gaming frequency, and level of experience were shown to significantly affect human trust and reliance behaviours [20, 21]. Similarly, uncontrolled environmental conditions, such as risk and uncertainty, have some implications on trust and its role in predicting the level of reliance on automation [22, 23].

Offline models for trust can be used by system designers to evaluate the performance of design alternatives at early stages of system development. Models that readily incorporate task-related human and environmental factors are helpful in estimating performance bounds under various environmental conditions and across different user categories. Furthermore, such models can be used to set human selection criteria as they enable the quantitative analysis of the role of individual factors on system performance.

In addition, as models serve as abstraction of real-world systems, the careful analysis of these models can lead to revealing potential modifications to system design that are likely to result in enhancing system performance. Below is a discussion on how computational models for trust can be utilised towards improved design of trust-aware systems.

5.1. Model leverage points

Trust models can be used to reveal leverage points in the system. According to Meadows, leverage points are " places within a complex system (a corporation, an economy, a living body, a city, an ecosystem) where a small shift in one thing can produce big changes in everything" [24, p. 1]. Leverage points are of particular significance to system designers as they are potentially the right places where wellfocused actions can lead to enduring improvements [25].

By closely analysing a system model, its leverage points can be identified by exploring positive or negative behaviours and by looking for the causes of these behaviours. For instance, model sensitivity analysis can be used to uncover some leverage points by identifying model parameters to which the outputs

are most sensitive. These parameters can represent automation characteristics, requirements on training, or criteria for human selection. Thus, rather than testing and evaluating the system under a wide range of possible settings, this process identifies potential changes to the system which will most likely result in considerable improvements. These potential changes can then be subjected to more thorough testing which can include actual human experiments. In this way, limited testing resources can be optimally allocated to investigating promising solutions.

Clare [3] used sensitivity analysis to identify important parameters in his system dynamics model which captures the relationship between trust, human interventions, workload, and system performance. His analysis revealed that initial trust and real-time human expectations of performance were the human-related parameters to which the performance was most sensitive. Based on these results, he proposed the use of positive/negative information framing about automation performance, as a way of priming humans to influence their level of initial trust. Besides, he investigated priming subjects' expectation in real-time using a graphical comparison that shows the difference between their own progress and the progress achieved by a group of other subjects who achieved high/low performance, to raise/lower their expectations.

However, there is a major challenge inherent in the mapping between model sensitive parameters and the actions needed to influence their corresponding qualitative variables. That is, while model analysis can help with the identification of leverage points, it does not say much, if anything, about what actions are required to utilise them. For instance, the model in [3] predicted that heightened human expectations would lead to a significant increase in system performance. Yet, it is beyond the model scope to predict whether the graphical comparison between subjects' progress and the progress of high performers is a suitable method of raising subjects' expectations. In fact, the results of the human experiments revealed that this method was not effective. The author explained that this method probably led to frustrating the subjects leading them to lower, rather than raising, their expectations. This example highlights the need for validating the effect of the proposed intervention on the leverage point of interest; which is by itself far from being straightforward.

Another related challenge lies in the quantification of the effect of a potential intervention on the qualitative parameter of interest. While the effect of the intervention might have been validated in other research studies, it can still be important to measure the size of this effect on the parameter of interest so that the model can generate a quantitative prediction of the net effect on system performance. For example, while it can be derived from past studies that carefully designed training programs have a significant effect on raising/lowering initial trust, the amount of change should be determined so that the adequacy of the intervention can be evaluated early on. The answer to these two questions (whether or not, and if so, by how much a given intervention affects the leverage point of interest) may need to be obtained through a separate model that is specially designed for such uses, or using actual human experiments. Figure 2 shows the process of identifying and utilising leverage points in a system.

In addition to identifying individual leverage points, sensitivity analysis for parameter combinations can be used to identify which model parameters should be changed together and in what direction to cause the maximum change in model output. Optimisation techniques are also useful to find candidate values for model parameters to optimise model output [26].

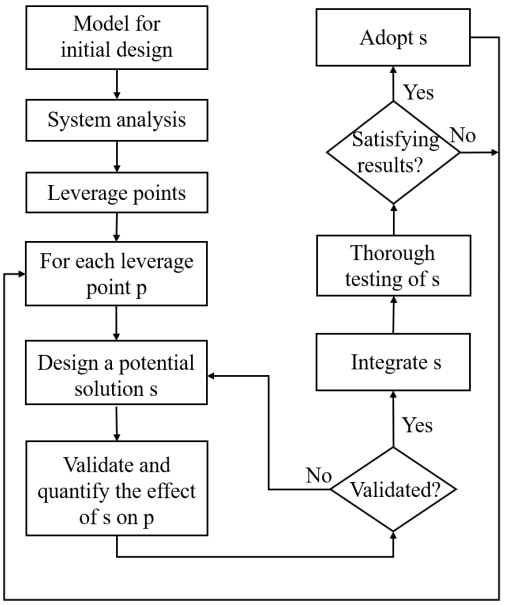

Figure 2. Identifying and utilising leverage points.

5.2. Customised improvements

The discussion in section 5.1 assumes that a design option has consistent effects among most people. However, there can be some interaction between individual differences and the proposed modifications, such that the modification may be effective only within some groups of people, while it may be ineffective or even detrimental among the others. Revisiting Clare's example of priming initial human

trust [3], the model predicted that low initial trust would benefit the performance as it would lead to increasing operators' interventions, which were needed for improved performance. However, the proposed solution was not found to be beneficial among the test subjects as there was no performance difference between those with heightened, lowered, and control initial trust. In addition, those with lowered initial trust had significantly more errors than subjects in the other groups. Clare suggested that this could be due to the high levels of workload accompanying the high frequency of interventions among subjects with lowered trust, which led to more errors. By considering only frequent video gamers, who had more spare mental capacity due to their high speed, the proposed solution was found to be effective such that subjects with lowered initial trust performed significantly better than those with heightened initial trust.

Another example is increased automation transparency which has been widely advocated by many researchers [2, 27, 28, 7] for its anticipated positive impacts on system performance through well-calibrated trust and proper reliance on automation. Nonetheless, mixed results have been reported in literature, such that some studies found that increased transparency improved proper reliance and system performance [29, 30], while others found that transparency led to lower efficiency without significant effect on other system performance metrics [31, 8]. While the implementation of transparency could be a source of this discrepancy, we also expect that individual differences in information processing and in cognitive capacity could have been at least partially responsible for these different findings. Thus, investigating individual differences that are known to affect human interaction with automation (e.g. relevant experience and frequency of playing video games) and incorporating them within the models, could be very useful for understanding the behaviours of different user groups, and hence designing customised modifications or training programs to improve their performance.

Besides individual differences, understanding how the system performs under different environmental conditions can be crucial for estimating its performance bounds. Closely related to trust are the concepts of risk and uncertainty without which trust may be considered irrelevant [32]. Thus, the inclusion of such uncontrolled contextual factors within trust models is important as it allows designers not only to have estimated measures of their impact on the dynamics of the interaction, but also to investigate the possibility of designing separate improvements for each condition.

6. Online Models in the Deployment of Trust-Aware Systems

Online models for trust can also be used to provide real-time assessments of the level of trust which can then be used to trigger appropriate responses. As discussed in section 3, online models for trust use the available real-time data to provide estimates of the level of trust. These estimates can be used in various ways to promote task performance by adapting automation behaviour, eliciting changes in human behaviour through adapting automation transparency, or adjusting the level of autonomy within the interaction.

6.1. Automation Behaviour Adaptation

Automation parameters can be hard-coded to produce an overall good performance across different scenarios. However, while performing the task, the human supervisor can appraise the situation and maintain a certain level of trust in the automation based on how well its behaviour given the specifics of the actual situation. Although the machine can adapt some of its parameters based on the state of the task, this adaptation is conditional upon its ability to sense relevant elements of the task, and determine how its behaviour should change accordingly. This can be very difficult in real environments which can be dynamic such that unmodelled events can take place.

Therefore, the ability to use the level of human trust in automation to assess how well the automation is performing and whether or not adaptation is required, is an appealing idea. In this way, an overall judgement of the performance can be obtained without requiring the human supervisor to explicitly communicate to the machine whether and how its behaviour should be adjusted. Xu and Dudek [18] used their trust model to estimate the real-time decline in the level of trust to identify whether behaviour adaptation is needed. In their work, a robot controller selects a set of parameters that defines the behavior of an unmanned aerial vehicle (UAV) in a visual navigation task in which the UAV is required to track the boundary of a given terrain. When a significant decrease in the user's level of trust is estimated, the robot controller adapts its parameters based on the amount of trust lost. For a small amount of estimated trust loss, the robot controller adjusts its parameters incrementally to finely tune its behaviour to improve the performance. However, if the the amount of trust lost is sufficiently large, the controller re-adjusts all its parameters by evaluating different configurations and modes of operation to determine which one will produce the most trusted behaviour. In that study, it

was found that the trust-driven adaptive behaviour led to a significant improvement in the performance as compared to the non-adaptive behaviour.

Another way in which automation can adapt its performance is to change the weight of different objectives of the performance in response to changes in the state of the task or to the subjective preferences of the human supervisor. For instance, in a task where both speed and accuracy are crucial for the performance, the speed-accuracy compromise adopted by the human supervisor can affect his/her selection of performance strategy. Thus, a supervisor who values accuracy more than speed will trust the machine that helps them achieve highly accurate operations than faster ones.

Floyd et al. [33] used a case-based reasoning approach to find a highly trusted behaviour that aligns with the preferences of the human operator. Their algorithm is based on the assumption that operators with similar preferences will react similarly towards the same robot's behaviours. Staring with an initial set of parameters, the robot calculates the trustworthiness of the resulting behaviour as a linear function of successful task completion rate and the rate of human intervention with robot operation. If the trustworthiness of the robot behaviour drops below a certain threshold, the robot adapts its behaviour by applying a random walk algorithm to change its parameters so that it can explore new behaviours. When a trusted behaviour is found, its parameters are recorded. Moreover, the history of behaviours, together with their trustworthiness values, that were used before reaching the trusted behaviour are recorded. This allows the robot to quickly find trusted behaviours for new operators with similar preferences rather than starting from scratch with every new human operator.

Trust-triggered behaviour adaptation is still in its early stage; and hence the factors affecting its success have not yet been fully investigated. One of these factors is that trust has inertia [9], which means that the effect of automation performance on trust is not instantaneous, but can occur gradually with some delays. That is, a decline in task performance at time t may result in a significant loss of trust, beyond the adaptation threshold, at time $t + \tau$. This means that the machine gets delayed information about its low performance aside from the time it needs to explore and find another set of parameters for a more trusted behaviour. The effects of these time delays in relation to the degree of dynamism of the environment should be investigated to understand how the adaptation threshold should be set accordingly.

Automation predictability is another important issue. While the adaptability of automation behaviour is a desirable capability, it can lead to perceived inconsistent behaviour as the automation may behave differently in the same situation each time the situation is repeated. As predictability is an important basis of trust [2], it is crucial to examine the effects of behaviour adaptation on the predictability of the automation and to investigate the need to correct for such effects, possibly through a transparent interface that keeps the human updated with these adaptations.

6.2. Transparency Adaptation

Automation transparency is a critical interface element as it enables trust calibration based on the understanding of how automation works and when it is more likely to fail. The challenge however is to determine the proper degree of transparency that maximises humans' situational awareness without overwhelming them with too much information that can considerably increase their workload. Towards this goal, Chen et al. [28] proposed the situation awareness based agent transparency (SAT) model, which is a conceptual model with three levels of transparency corresponding to the three levels in Endsley's model for situational awareness [34]. The SAT model is aimed at calibrating trust in real-time while maintaining the desired level of situational awareness. Chen et al. suggested that the proper level of transparency should be determined based on the state of the task to increase the effectiveness of task performance.

Akash et al. [16] proposed a partially observable Markov decision process (POMDP) model for the real-time inference of human trust and workload in robot-assisted reconnaissance missions. Their work presents the first attempt to adapt transparency based on the estimated level of trust. They defined the reward function in their POMDP model in terms of the level of trust, workload, and task performance; and they used transparency as the feedback to maximise this reward function. They found that their proposed algorithm for adapting the level of transparency led to significant positive effects on human workload as well as mission success and efficiency.

6.3. Flexible Autonomy

Another way in which systems can be adapted based on the estimated level of trust is by sliding the level of autonomy in shared-control tasks. Past studies investigated different ways of performing this adaptation capability by: allocating the task to either the human or the machine, changing the level of automation proactivity, or by adjusting the weights of the manual and automation inputs.

Wang et al. [12] proposed a model for the mutual

trust between a human and a robot based on their individual performance and the overall fault rate. Human performance is calculated in terms of human utilization which is a function of the portion of time the human is manually controlling the robot and the difficulty of the task. Meanwhile, robot performance is modelled to decrease overtime when the human is not manually operating it and to increase overtime otherwise. According to the level of estimated mutual trust, task control at a given time step can be exclusively allocated to the robot or the human.

Sadrfaridpour et al. [35] proposed a real-time model for trust estimation in human-robot collaboration in manipulation tasks. In that work, the human and the robot share the task of impedance control by applying force to the manipulator to move it to the right position. The impedance control mode used by the robot is selected based on the level of trust. When low levels of trust are estimated, the robot operates in a reactive mode such that the human performs the motion planning problem which places more workload demands on the human. On the other hand, when high levels of trust are predicted, the robot activates its proactive mode in which it estimates the human desired motion and acts accordingly to share the effort of the motion planning with the human.

Saeidi et al. [36] used the estimated level of human trust in automation to integrate the manual and autonomous control inputs in the teleoperation of mobile robots, such that the control of the robots is continuously shared between the human and the automation. When high/low levels of trust are estimated, the interface gives high/low weights for the machine control inputs, respectively. Saeidi et al. also used another model for calculating robot-to-human trust based on human performance. When the calculated trustworthiness of the human drops, the robot communicates this information to the human through haptic feedback.

7. Towards Effective Trust Models

Based on the discussion so far, it is evident that computational models for trust can be used in a variety of ways to deliver trust-aware systems. We cited some examples from literature to consolidate the potential uses of trust models and to give some evidence on their promising effects on system performance. Nevertheless, utilising trust models within the design and deployment of trust-aware systems is still a largely unexplored area. This section sheds the light on some aspects of quantitative trust models that enhance their effectiveness and usability by system designers.

Firstly, despite the great interest in the concept of human trust in automation, it is worth recalling that trust is not an end goal by itself, rather enhanced system performance that results from proper reliance is what really matters [37]. Hence, a model that is solely intended to estimate the level of trust without being able to give accurate predictions on human reliance can be of limited value to system designers. The usability of trust models will be largely dependent on how far they capture the causes (e.g performance) and effects (e.g reliance rate) of trust as well as the implications on overall performance. Furthermore, depending on the intended use of the model, it may need to include additional environmental and individual factors.

Models that aim to help designers with the evaluation of design options under different conditions will need to have representation of these conditions within the model boundaries. For instance, the level of task-associated risk was found to moderate the effect of trust on reliance such that at high risk situations the rate of reliance decreases although trust remains unchanged [22, 23]. Thus, a model for a system that is expected to encounter risky conditions should include the effects of risk on user trust and reliance. The inclusion of individual factors can also be warranted by the expected variability among system users. Systems that target a wide sector of users (e.g, autonomous vehicles) need their models to include the effects of relevant individual factors like age and relevant experience [38]. Modelling such individual and environmental factors allows the designers not only to investigate the performance under different conditions, but also to investigate the possibility of having customised designs for these conditions.

The increasing interest in studying transparency and the notable trend of delivering transparent automation suggest that models incorporating transparency may prove useful in giving accurate predictions both for the rates of reliance and for its appropriateness with respect to the state of the system. This will enable a closer investigation of the frequency of undesirable reliance behaviours (incorrect reliance and incorrect rejection) such that preventive or corrective actions can be considered by designers.

Finally, the possible effects of potential system modifications or real-time adaptations on human workload and situational awareness should be included within the model to avoid undesirable consequences on performance through unmodelled factors. For example, although transparency can benefit trust and situational awareness, it can equally have negative effects on workload. Also, a highly trusted machine can operate at a higher level of autonomy to mitigate workload but this may hurt situational awareness.

8. Conclusion and Future Work

In this work, we investigated the use of computational models for trust within two development stages, namely design and deployment. We showed how offline trust models can be used by designers to evaluate performance under different conditions and to identify possible improvements/interventions. We also presented promising uses of online models to trigger adaptation based on the estimated level of trust. However, for the models of trust to reach their potential, they should be adequately representative of the actual systems by capturing the key factors that affect trust and reliance, as discussed in section 7. Our work presents an interdisciplinary overview of literature to answer the three research questions posed in the paper. Past studies, considered in this work, were sampled from the literature on developing, evaluating, and using computational models for trust. The studies were selected to represent the different ways in which such models can be used by system designers. Thus, our work paves the way towards the development of a framework for trust-aware human-automation interaction.

Past studies that utilised trust models for real-time adaptation demonstrated improved system performance. Nevertheless, it can be noted that most of these studies were mainly concerned with the loss of trust and proposed adaptations to correct for such a loss. As over-trust can also be detrimental to the performance [2], we believe that over-trust triggered adaptations can be equally important to the safe and effective system operation. These adaptations may include increased transparency to reveal system weaknesses, but should be done carefully to avoid its counter-effects of drastically losing trust.

Systems' ability to adapt to changes in the environment is a desirable capability as it ensures a system's flexibility and continued usability in dynamic environments. However, adaptation may lead to decreased predictability of automation behaviour and hence can hurt human trust in it. Researchers believe that user trust in adaptive systems is a must without which the users are likely to abandon these systems [39]. This raises a few questions regarding the design of trust-aware automation with adaptation capabilities. First, how can we use trust calibration mechanisms, e.g transparency, to mitigate the side effects of adaptation on trust? Second, can adaptation that is driven by trust, possibly among other variables, be designed to avoid the negative side effects on future trust values? And finally, how the estimated level of trust can be used to correctly trigger the adaptation? That is, a loss in trust which is caused by performance drop might need different adaptation than a loss in trust caused by unpredictability.

While the focus of this paper is on trust, many of the discussed concepts can be applied to other human factors, such as workload and situational awareness. In fact, a holistic model that combines the interactions among these three factors can be useful for representing situations where these factors are relevant. This will enable developing suitable interventions or adaptations that take into consideration the combined effect of these factors on system performance.

Acknowledgement

This work was funded by the Australian Research Council Discovery Grant number DP160102037 and UNSW-Canberra.

References

[1] H. A. Abbass, "Social integration of artificial intelligence: Functions, automation allocation logic and human-autonomy trust," *Cognitive Computation*, vol. 11, no. 2, pp. 1–13, 2019.

[2] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, no. 1, pp. 50–80, 2004.

[3] A. S. Clare, *Modeling real-time human-automation collaborative scheduling of unmanned vehicles*. PhD thesis, Massachusetts Inst of Tech Cambridge Dept of Aeronautics and Astronautics, 2013.

[4] S. Levin, "Tesla fatal crash: 'autopilot' mode sped up car before driver killed, report finds," *The Guardian*, June 2018.

[5] The Tesla Team, "An update on last week's accident," March 2018.

[6] K. A. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," *Human Factors*, vol. 57, no. 3, pp. 407–434, 2015.

[7] S. Ososky, T. Sanders, F. Jentsch, P. Hancock, and J. Y. Chen, "Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems," in *Unmanned Systems Technology XVI*, vol. 9084, p. 90840E, International Society for Optics and Photonics, 2014.

[8] T. Helldin, U. Ohlander, G. Falkman, and M. Riveiro, "Transparency of automated combat classification," in *International Conference on Engineering Psychology and Cognitive Ergonomics*, pp. 22–33, Springer, 2014.

[9] J. Gao and J. D. Lee, "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*, vol. 36, pp. 943–959, Sep 2006.

[10] F. Gao, A. S. Clare, J. C. Macbeth, and M. Cummings, "Modeling the impact of operator trust on performance in multiple robot control," AAAI, 2013.

[11] A. Hussein, S. Elsawah, and H. Abbass, "A system dynamics model for human trust in automation under speed and accuracy requirements," in *Proceedings of*

the Human Factors and Ergonomics Society Annual Meeting, 2019.

[12] Y. Wang, Z. Shi, C. Wang, and F. Zhang, "Human-robot mutual trust in (semi)autonomous underwater robots," in *Studies in Computational Intelligence*, pp. 115–137, Springer Berlin Heidelberg, 2014.

[13] A. Xu and G. Dudek, "Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pp. 221–228, ACM, 2015.

[14] X. J. Yang, V. V. Unhelkar, K. Li, and J. A. Shah, "Evaluating effects of user experience and system transparency on trust in automation," in *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, pp. 408–416, IEEE, 2017.

[15] C. Nam, P. Walker, M. Lewis, and K. Sycara, "Predicting trust in human control of swarms via inverse reinforcement learning," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, IEEE, Aug 2017.

[16] K. Akash, T. Reid, and N. Jain, "Improving human-machine collaboration through transparency-based feedback–part ii: Control design and synthesis," *IFAC-PapersOnLine*, vol. 51, no. 34, pp. 322–328, 2019.

[17] M. Hoogendoorn, S. W. Jaffry, and J. Treur, "Cognitive and neural modeling of dynamics of trust in competitive trustees," *Cognitive Systems Research*, vol. 14, no. 1, pp. 60–83, 2012.

[18] A. Xu and G. Dudek, "Trust-driven interactive visual navigation for autonomous robots," in *2012 IEEE International Conference on Robotics and Automation*, IEEE, May 2012.

[19] J. Y. Chen and M. J. Barnes, "Human–agent teaming for multirobot control: A review of human factors issues," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 1, pp. 13–29, 2014.

[20] M. Dikmen and C. Burns, "Trust in autonomous vehicles: The case of tesla autopilot and summon," in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE, Oct 2017.

[21] S. M. Merritt and D. R. Ilgen, "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*, vol. 50, pp. 194–210, Apr 2008.

[22] K. Satterfield, C. Baldwin, E. de Visser, and T. Shaw, "The influence of risky conditions in trust in autonomous systems," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 61, pp. 324–328, Sep 2017.

[23] G. Sadler, H. Battiste, N. Ho, L. Hoffmann, W. Johnson, R. Shively, J. Lyons, and D. Smith, "Effects of transparency on pilot trust and agreement in the autonomous constrained flight planner," in *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*, IEEE, Sep 2016.

[24] D. Meadows, "Leverage points-places to intervene in a system," 1999.

[25] P. M. Senge, "The fifth discipline: The art and practice of the learning organization (rev. ed.)," *New York, NY: Currency Doubleday*, 2006.

[26] J. P. Kleijnen, "Sensitivity analysis and optimization of system dynamics models: regression analysis and statistical design of experiments," *System Dynamics Review*, vol. 11, no. 4, pp. 275–288, 1995.

[27] J. B. Lyons, "Being transparent about transparency," in *AAAI Spring Symposium*, 2013.

[28] J. Y. Chen, K. Procci, M. Boyce, J. Wright, A. Garcia, and M. Barnes, "Situation awareness-based agent transparency," tech. rep., Army Research Lab Aberdeen Proving Ground MD Human Research and Engineering Directorate, 2014.

[29] J. E. Mercado, M. A. Rupp, J. Y. Chen, M. J. Barnes, D. Barber, and K. Procci, "Intelligent agent transparency in human–agent teaming for multi-uxv management," *Human Factors*, vol. 58, no. 3, pp. 401–415, 2016.

[30] N. Wang, D. V. Pynadath, and S. G. Hill, "Trust calibration within a human-robot team: Comparing automatically generated explanations," in *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, pp. 109–116, IEEE Press, 2016.

[31] T. Helldin, *Transparency for Future Semi-Automated Systems: Effects of transparency on operator performance, workload and trust*. PhD thesis, Örebro Universitet, 2014.

[32] A. C. Wicks, S. L. Berman, and T. M. Jones, "The structure of optimal trust: Moral and strategic implications," *The Academy of Management Review*, vol. 24, p. 99, Jan 1999.

[33] M. W. Floyd, M. Drinkwater, and D. W. Aha, "Learning trustworthy behaviors using an inverse trust metric," in *Robust Intelligence and Trust in Autonomous Systems*, pp. 33–53, Springer US, 2016.

[34] M. R. Endsley, "Situation awareness global assessment technique (sagat)," in *Aerospace and Electronics Conference, 1988. NAECON 1988., Proceedings of the IEEE 1988 National*, pp. 789–795, IEEE, 1988.

[35] B. Sadrfaridpour, M. F. Mahani, Z. Liao, and Y. Wang, "Trust-based impedance control strategy for human-robot cooperative manipulation," in *ASME 2018 Dynamic Systems and Control Conference*, pp. V001T04A015–V001T04A015, American Society of Mechanical Engineers, Sep 2018.

[36] H. Saeidi, F. McLane, B. Sadrfaidpour, E. Sand, S. Fu, J. Rodriguez, J. R. Wagner, and Y. Wang, "Trust-based mixed-initiative teleoperation of mobile robots," in *2016 American Control Conference (ACC)*, IEEE, Jul 2016.

[37] C. F. Rusnock, M. E. Miller, and J. M. Bindewald, "Observations on trust, reliance, and performance measurement in human-automation team assessment," in *IIE Annual Conference. Proceedings*, pp. 368–373, Institute of Industrial and Systems Engineers (IISE), 2017.

[38] M. Dikmen and C. Burns, "Trust in autonomous vehicles: the case of tesla autopilot and summon," in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 1093–1098, IEEE, 2017.

[39] D. S. Lange, "Trust of, in, and among adaptive systems," in *Monterey Workshop*, pp. 193–205, Springer, 2010.