Trust in Automation

A Literature Review

Bronwyn French\textsuperscript{1}, Andreas Duenser\textsuperscript{2}, Andrew Heathcote\textsuperscript{1}

EP184082

22.05.2018

\textsuperscript{1} School of Medicine, Division of Psychology, University of Tasmania

\textsuperscript{2} Data61, CSIRO
Increasingly complex autonomous systems require the human operator to appropriately calibrate their trust in the automation in order to achieve performance and safety goals. Although human-automation trust has been shown to be similar in many respects to human-human trust, precisely defining and measuring trust in automation has proved to be one of the greatest challenges facing research in this area. This report reviews literature pertaining to trust in automated systems to provide an integrated summary of the major theoretical and empirical work in the field to date.
## Contents

Abstract .......................................................................................................................... 2

Acknowledgments ............................................................................................................ 5

1 Introduction .................................................................................................................. 6

2 Automation ..................................................................................................................... 8
   2.1 Definitions of automation ......................................................................................... 8
   2.2 Taxonomies of automation levels ........................................................................... 9
   2.3 Theoretical models of automation ......................................................................... 15
   2.4 Appropriate use of automation: use, misuse, disuse, abuse ................................. 17

3 Trust .............................................................................................................................. 19

4 Trust in automation ....................................................................................................... 21
   4.1 The importance of examining trust in the use of automation ................................. 21
   4.2 Trust in people generalised to trust in automation ................................................. 21
   4.3 Defining trust in automation .................................................................................. 22
   4.4 Theoretical basis of trust formation ..................................................................... 24
   4.5 Appropriate calibration of trust in automation ....................................................... 26

5 Models of trust in automation ....................................................................................... 29
   5.1 Early models ........................................................................................................... 29
   5.2 Recent and current models ...................................................................................... 32
   5.3 Models of specific facets of trust in automation ..................................................... 38
   5.4 Models of specific types of automation ................................................................ 40

6 Factors influencing trust in automation ..................................................................... 42
   6.1 Operator factors ..................................................................................................... 42
   6.2 Automation factors ............................................................................................... 44
   6.3 Environmental factors ......................................................................................... 46
   6.4 Learned trust – Dynamic trust changes over time ................................................. 47

7 Measuring trust in automation ..................................................................................... 49
   7.1 Issues measuring trust in automation ................................................................... 49
   7.2 Current trust in automation measures .................................................................. 51
Acknowledgments

This report was completed as part of the Automation, Trust and Workload Collaborative Research Project (CRP)
1 Introduction

Automation is changing the world in which we live. Rapid advances in automation technologies, including autonomous vehicles, robotics, autonomous web-based systems, and user experience frameworks and decision aids, are dramatically altering jobs and impacting upon many areas of daily life. A recent report by McKinsey Global Institute (2017) estimates that by 2030 up to 30% of work activities could be displaced by automation, with up to 14% of the global workforce likely to need to transition to new occupations. Most occupations will have their work practices altered as automated systems increasingly work alongside humans. Between 2000 and 2015, an average of 6% of tasks performed by Australian workers were automated (AlphaBeta, 2017).

Automation encompasses technology that actively selects and transforms data, makes decisions, or controls processes (Lee & See, 2004). Automation provides operational speed and efficiencies, increased productivity, improved reliability and sustainability, and the ability to complete tasks in difficult or unsafe environments. However much of this advancement is led by what is technically possible or economically expedient. For automation to work most beneficially for humans, consideration of the human role in working with autonomous systems is vital (Lee, 2008). Research has demonstrated that automation does not merely replace human activities, but rather alters it in ways which pose different demands upon the human operator (Parasuraman, Sheridan, & Wickens, 2000). Automation extends the capacity of people to achieve tasks which might otherwise not be possible, but this can only be achieved if the automation’s design considers the characteristic of the joint system of human and automation combined (Vagia, Transeth, & Fjerdingen, 2016).

As autonomous systems become more complex, the ability of human operators to understand the system becomes diminished. Trust has been identified as a key factor influencing reliance on automation (Lee & See, 2004). In particular, trust is significant in determining the willingness of a human operator to rely on automation in situations of uncertainty (Lee & See, 2004). Irrespective of the robustness of an autonomous system, it is likely that it will in some instances fall short of expectations. Therefore, trust in the system must be appropriately calibrated to the actual system performance (Muir, 1994). Over-trust may lead the human operator to depend upon the automation to perform outside its parameters or when faulty. Under-trust may lead to disuse of the automation, resulting in excessive operator workload and diminished system performance (Lee & See, 2004). Thus, facilitating appropriate levels of trust in automation is central to maximising performance and safety of human-automation teams (Hoff & Bashir, 2015).
This review provides a summary of scientific literature on trust in automation. Definitions, taxonomies, and theoretical models of automation are presented, and commonly used definitions of use, misuse, and disuse of automation are discussed.

Next, the review considers trust, first in the context of interpersonal or organisational relationship, and then relates that to human-automation trust and why it is an important field of study. Common definitions of trust used in the automation literature are presented, and the most commonly used theoretical framework for the foundations of trust is outlined. Relationships between trust and similar (possibly confounding) terms are presented and defined as far as possible within the limitations of the trust in automation literature. Appropriate calibration of trust is discussed, referring back to the descriptions of automation use, misuse, and disuse.

Significant past and current theoretical models of trust in automation are then presented and outlined. The review then considers factors affecting levels of trust in automation, dividing them into properties of the user, the situation, and the automation. Finally, the difficult topic of measuring trust is addressed.
2 Automation

2.1 Definitions of automation

Automation is most simply defined as the process by which a machine carries out a function previously completed by a human (Parasuraman & Riley, 1997). As automation capabilities have increased, this definition has expanded to include technology-executed tasks which augment and compensate human performance capabilities, and tasks which humans are unable to perform (Currie, 2007). This has led to more general definitions, such as “the use of technologies to improve processes and outcomes with substantially reduced reliance on human involvement” (Australian Industry and Skills Committee, 2017).

Lee and See’s (2004) widely-used definition of automation summarises the processes performed by the automation, defining automation as “technology that actively selects data, transforms information, makes decisions, or controls processes”. Other definitions which focus on functions are similar in their conceptualisation. For example, Sheridan (2002) defines automation as “the mechanisation and integration of the sensing of environmental variables; data processing and decision making; and mechanical action” (including ‘information action’). However, in research concerned with human performance in automated systems, automation does not refer simply to any technological innovation (Parasuraman, Sheridan, & Wickens, 2000). For example, improvements in battery technology do not necessarily constitute automation. Instead, automation in human factors research is most often considered as a system or device that “accomplishes (partially or fully) a function that was previously, or conceivably could be, carried out (partially or fully) by a human operator” (Parasuraman et al., 2000; p 287). However Bradshaw, Hoffman, Johnson, and Woods (2013) go further, and caution against thinking of autonomy as a discrete property of a work system or a particular kind of technology. Instead, they argue that autonomy is an “idealised characterisation of observed or anticipated interactions between the machine, the work to be accomplished, and the situation” (p 4).

What is defined as automation changes over time. When a function is reallocated fully and permanently to automation for performance and is completed with a consistently high level of success, the function tends to be viewed simply as machine operation, rather than automation (Parasuraman & Riley, 1997).
2.1.1 Automation, autonomy, autonomous systems

Within the literature on the subject, the terms automation, autonomy, and autonomous systems are often used interchangeably. However, a number of authors argue for differences between the meanings of the terms. Hancock (2017) and Endsley (2017) both argue that automation is limited in capability, intended to complete a specific set of processes in order to achieve one of a set of predetermined outcomes. Both authors consider autonomous systems to be more advanced automation which can learn and evolve, and as a result can change functional capacities and actions over time. Hancock (2017), therefore, proposes that the growth of autonomy rests upon increasing levels of automation. While some facets of automation evolve into autonomy, for other tasks and requirements automation will continue to be sufficient.

2.2 Taxonomies of automation levels

Technological developments have made it possible for virtually all human-machine systems to include aspects of automation. Deciding whether a particular function is better accomplished by a person or technology, and to what extent a function should be automated for optimal performance, has been a core aspect of human-machine systems research and design (de Winter & Dodou, 2014). A number of taxonomies which categorise the allocation of functions to humans or automation have been developed. A recent literature review comparing twelve of the main different taxonomies used in automation literature concluded that there was no single “best” taxonomy which applied across all possible automation applications (Vagia, Transeht, & Fjerdingen, 2016). This review therefore outlines only the taxonomies most commonly used in the trust in automation literature, and then considers current and future taxonomic directions.

2.2.1 The Fitts List

Originally function allocation was conceived in all-or-nothing terms and formalised in frameworks such as the Fitts list (1951), which was created to support air-navigation and air-traffic-control human engineering research. The list comprised a series of statements regarding the relative abilities and strengths of humans and machines (see Table 1). Automation was considered appropriate for any function in which the machine surpassed human ability (e.g., ability to perform repetitive, routine tasks); for functions in which a human surpassed machine (e.g., ability to improvise), it made no sense to attempt to automate. The Fitts list has been widely used in function allocation research, and is still considered adequate by many as a starting point (de Winter & Dodou, 2014). However, rather than an all-or-nothing phenomenon, automation is now seen as a more complex dimensional integration of human and technology, and the Fitts list is generally considered inferior to frameworks which consider more dynamically the processes involved in human and automation interaction (Dekker & Woods, 2002).
Table 1. The original Fitts list (from Fitts 1951, p. 10)

<table>
<thead>
<tr>
<th>Humans appear to surpass present-day machines in respect to the following:</th>
<th>Present-day machines appear to surpass humans in respect to the following:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Ability to detect a small amount of visual or acoustic energy</td>
<td>1. Ability to respond quickly to control signals and to apply great force smoothly and precisely</td>
</tr>
<tr>
<td>2. Ability to perceive patterns of light or sound</td>
<td>2. Ability to perform repetitive, routine tasks</td>
</tr>
<tr>
<td>3. Ability to improvise and use flexible procedures</td>
<td>3. Ability to store information briefly and then to erase it completely</td>
</tr>
<tr>
<td>4. Ability to store very large amounts of information for long periods and to recall relevant facts at the appropriate time</td>
<td>4. Ability to reason deductively, including computational ability</td>
</tr>
<tr>
<td>5. Ability to reason inductively</td>
<td>5. Ability to handle highly complex operations, i.e. to do many different things at once</td>
</tr>
<tr>
<td>6. Ability to exercise judgement</td>
<td></td>
</tr>
</tbody>
</table>

2.2.2 Sheridan and Verplank: Levels of automation

Sheridan and Verplank (1978) were the first to create a formal taxonomy of automation levels to outline the task responsibility distribution between human and automation. Formulated to consider human and computer decision-making control of teleoperated submersible vehicles, they defined 10 levels of automation, from fully manual to fully autonomous, in which each level increases the amount of authority delegated to the automation, and decreases the operator involvement in decision making (see Table 2).

Table 2. Sheridan and Verplank’s (1978) levels of automation in man-computer decision-making

<table>
<thead>
<tr>
<th>Level</th>
<th>Description of interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Human does the whole job up to the point of turning it over to the computer to implement</td>
</tr>
<tr>
<td>2</td>
<td>Computer helps by determining the options</td>
</tr>
<tr>
<td>3</td>
<td>Computer helps determine options and suggests one, which human need not follow</td>
</tr>
<tr>
<td>4</td>
<td>Computer selects action and human may or may not do it</td>
</tr>
<tr>
<td>5</td>
<td>Computer selects action and implements it if human approves</td>
</tr>
<tr>
<td>6</td>
<td>Computer selects action, informs human in plenty of time to stop it</td>
</tr>
<tr>
<td>7</td>
<td>Computer does whole job and necessarily tells human what it did</td>
</tr>
<tr>
<td>8</td>
<td>Computer does whole job and tells human what it did only if human explicitly asks</td>
</tr>
<tr>
<td>9</td>
<td>Computer does whole job and tells human what it did and it, the computer, decides he</td>
</tr>
<tr>
<td>10</td>
<td>Computer does whole job if it decides it should be done, and if so tells human, if it decides</td>
</tr>
</tbody>
</table>

This approach to automation is one of the most widely cited, and is the basis of many subsequent taxonomies which focus on changes in the level of approval the human operator exerts over the automated system’s decisions (e.g., Endsley, 1987; Endsley & Kiris, 1995; Endsley, 1999; Draper, 1995; Proud, Hart, & Mrozinski, 2003).

In a similar taxonomy, Riley (1989) additionally included the intelligence of the automation as well as the level of autonomy. Variance in these two dimensions resulted in 12 levels of automation.
2.2.3 Parasuraman, Sheridan, and Wickens: Types and levels of human interaction with automation

Parasuraman, Sheridan, and Wickens (2000) created a framework which could be used to provide an objective basis for making function allocation choices, intended to evaluate the human’s interaction with the automation, rather than simply outlining the capabilities of the automation. They outlined four classes or types of automation function, derived from models of human information processing:

1. information acquisition
2. information analysis
3. decision and action selection
4. action implementation

The framework proposes that automation within each of the four types can be applied along a continuum, from fully manual (level 1) to fully automatic (level 10), in a similar manner to Sheridan and Verplank’s (1978) levels of automation. At levels 1 to 5 the human has overall control of a task, but increasingly utilises automation for assistance, while at levels 6 to 10 the automation completes tasks independently, providing human operators with decreasing feedback at each level (Hoff & Bashir, 2015). A particular automated system can involve a combination, or all, of these automation types at different automation levels.

Information acquisition automation involves the sensing and recording of input data (Parasuraman et al., 2000). At lower levels, this may consist of functions such as automatically moving sensors in order to observe an environment. At a moderate level of automation, the observation information may be organised according to given criteria. At higher levels of automation, the information may additionally be filtered, with only certain information displayed to the human operator.

Information analysis automation involves information manipulation or inference processes. At low automation levels, this may consist of an algorithm applied to information to extrapolate over time in order to make predictions or display trends. At higher automation levels, several input information streams may be combined and summarised to provide context-dependent summaries.

Decision and action selection automation involves augmenting or replacing human decision making. Levels of automation follow those described in Sheridan and Verplank’s (1978) levels above, with systems automated at lower levels suggesting action options for the human operator to choose from, and those at higher levels selecting the actions independently.

Action implementation automation involves allocation of the carrying out of the action chosen. The specific level of automation is defined by the relative amount of human versus technological activity in completing the action.

An additional, less used, class of automation has been proposed: automation that exists solely for the purpose of managing other automation (Hoff & Bashir, 2015). This may be less used as it has been viewed as simply a particular case of automation consisting of a combination of the other classes (Adams, Bruyn, Houde, & Angelopoulos, 2003).
Higher degrees of automation may be achieved both by later types (e.g., automated decision aid rather than information analysis) and later levels (e.g., completing a chosen task unless vetoed, versus offering several choices to the human operator). Wickens, Li, Santamaria, Sebok, and Sarter (2010) refer to this combination of types and levels as degrees of automation (see Figure 1). They note that, in general, as the degree of automation increases, routine automation performance is improved while performance under failure declines. However, varying the degree of automation by stage may impact differently to varying it by level.

![Figure 1](image)

**Figure 1.** Degrees of automation (from Wickens et al., 2010).

While Sheridan and Verplank’s (1978) levels of automation were used in the framework to specify automation levels for decision automation, Parasuraman et al. (2000) did not explicitly specify levels for information automation, as they considered that such specifications would be superseded by technological developments in information integration and presentation methods. Much of the research subsequently utilising this taxonomy similarly refers more generally to ‘higher’ or ‘lower’ levels of automation rather than specific, numbered levels.

A meta-analysis of studies examining the human performance consequences of degrees of automation found benefits for routine performance and workload as degrees of automation increase, but a negative impact of higher degrees of automation when systems fail, and a negative impact on situation awareness and manual skills (Onnasch, Wickens, Li, & Manzey, 2014). They found that the negative consequences of automation seem most probable when the level of automation moves from information analysis to action selection.

### 2.2.4 Static and adaptive automation

A system of automation in which the levels of automation remain fixed once designed and implemented is referred to as static automation. Automation which has the ability to change level or type of automation during operation when initiated by either the automation or the operator is known as adaptive automation (Parasuraman, Bahri, Deaton, Morisson, & Barnes, 1992). Adaptive automation allows for the dynamic allocation of function between human and machine, with the level or type of automation flexibly altered to adapt to changes in workload, situational, or other peripheral factors. Ability to adapt automation level allows the human supervisor to take greater control in case of automation faults or errors; alternatively, it allows the automation to move to a higher level to compensate for occasions of high human operator workloads (Vagia, Transeht, &
Fjerdingen, 2016). This allows for performance level to be maintained, and it has the additional benefit of enabling the human operator to be more consistently engaged, resulting in better situation awareness (Miller & Parasuraman, 2007).

2.2.5 Current and future directions in taxonomy of automation levels

Levels of autonomy can provide a useful concept for research, design, and training purposes, but they can also prove problematic, as such descriptions assume a linear hierarchy that often does not exist in practice (Abbott, 2015). In many of the more complex autonomous systems, such as autoflight systems on aeroplanes, the various features of the system may be selected in a variety of combinations that cannot be ascribed to a simple hierarchical description. In 2012, the US Department of Defense [sic] concluded that studies based on levels of autonomy are counter-productive to the autonomy design process, as they focus attention upon the technology at the expense of the collaboration between automation and human operator, and additionally do not support flexibility of automation levels within a single system, leading to designs which deliver explicit functions rather than provide general resilient capability (Murphy & Shields, 2012). It recommended the cessation of the use of conceptual frameworks based on levels of autonomy, to be replaced by autonomous systems reference frameworks that focus upon the allocation of cognitive functions and responsibilities between human and automation.

This view is supported by Bradshaw, Hoffman, Johnson, and Woods (2013), who argue that the notion of levels of automation are problematic for additional reasons:

1. Levels reinforce the erroneous notion that automation and human activities are interchangeable, and that substituting one for the other has no effect on the system’s operation.
2. It is not always apparent whether a given action is ‘higher’ or lower’ than another action; additionally, in a specific situation, an automated system may be performing high and low actions simultaneously.
3. Level of autonomy is relative to the task, goals, and context
4. Levels of autonomy promote reductive thinking, for example, considering simultaneous activity as sequential activities
5. The concept of levels of autonomy is insufficient to meet current and future challenges, for example when considering human-machine teamwork
6. The concept of levels of autonomy is not “human-centred”
7. The levels provide insufficient guidance to designers of automated systems

However, Kaber (2018) disagrees, arguing that discarding this framework makes little sense unless better alternatives can be provided for use by the wider design community. He addresses concerns with the framework of steps and levels of automation, concluding by proposing that a better approach in order to support broader automation design practice would be to incrementally advance the levels of automation approach, such as by using empirical studies to make taxonomies more descriptive. This view is supported by other prominent automation researchers such as Endsley (2018). Lee (2018) sees merit in both arguments but proposes that considering automation as a network of interconnected components may ultimately prove more beneficial than simply
refining function allocation frameworks. This perspective is intended to counter the propensity to
decompose complex systems into components which are subsequently erroneously assumed to be
independent.

The Level of Human Control Abstraction (LHCA) framework of automation is one framework which
has been proposed in response to Murphy and Shield’s (2012) report to consider vehicle control
automation (Johnson, Miller, Rusnock, & Jacques, 2017). It aims to provide a dynamic and human-
centric perspective by focusing on the actions and responsibilities required, at any moment in time,
of a human operator when controlling an automated system. Through this, it aims to provide a
mechanism for understanding the transfer of cognitive tasks from human operators to automated
systems. The framework considers automation in terms of the control tasks performed by the
automation and the level of detail of decisions made by the human operator. It proposes five levels
of control: direct control, augmented control, parametric control, goal-oriented control, and
mission-capable control (see Table 3). The framework suggests that, as the level of control detail is
reduced (from direct to mission-capable), the level of human attention and workload is reduced.
However, the level of abstraction is not considered a static attribute of the system. The LHCA is
determined at any moment in time by the most detailed control input given by the operator at that
moment, and a given level might correspond to a system state rather than the overall system. This
allows for consideration of adaptable or adaptive automation.

Table 3. Level of Human Control Abstraction framework (adapted from Johnson et al., 2017)

<table>
<thead>
<tr>
<th>Level of Human Control Abstraction</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Direct control</td>
<td>The operator controls every aspect of the system’s operation</td>
<td>The simplest possible fixed wing remote control aircraft</td>
</tr>
<tr>
<td>2. Augmented control</td>
<td>The operator gives control inputs commanding desired actions, the system then makes final determinations about control surface positions or motor power</td>
<td>A multi-rotor unmanned aerial system</td>
</tr>
<tr>
<td>3. Parametric control</td>
<td>The operator inputs desired parameters that the system should meet, the system then uses on-board sensors and control algorithms to meet those parameters</td>
<td>A commercial airliner with autopilot activated</td>
</tr>
<tr>
<td>4. Goal-oriented control</td>
<td>The operator inputs desired goals the system should meet, the system then makes all required decisions to meet those goals</td>
<td>A multi-rotor unmanned aerial system executing a goal of keeping a specified target centred in the camera’s field of view, avoiding obstacle collisions, and following the target until the command is disengaged by the human operator</td>
</tr>
<tr>
<td>5. Mission-capable control</td>
<td>The operator enters pre-launch mission goals at a level of detail which, when combined with standard operating procedures and</td>
<td>An autonomous car that travels to a desired location with no inputs from the driver other than the desired location</td>
</tr>
</tbody>
</table>
Taxonomies which focus on the interdependency between human and automation (Johnson, Bradshaw, & Feltovich, 2017) and on human-automation teams (Joe, O’Hara, Hugo, & Oxstrand, 2015) have also been developed. It remains unclear the extent to which these or similar frameworks might be adopted in automation research. In current research, Parasuraman, Sheridan, and Wickens’ (2000) types and levels of autonomy framework remains the predominant taxonomy. Due to the extensive use of levels of autonomy in engineering and design environments, moves away from the taxonomy in human factors research have the potential to lead to lack of clarity in communicating ideas.

Some fields of autonomous research have developed taxonomies specific to their area. In particular, the Society of Automotive Engineers (SAE) International’s (2014) J3016 standard defines a five-level taxonomy for autonomous vehicles, with specific performance-based criteria for each level (see Figure 2). The level of automation is defined in reference to the role played by the driver, the driving automation system, and other vehicle systems and components. This taxonomy is widely used in current vehicle autonomy research (e.g., Marinik, Bishop, Fitchett, Morgan, Trimble, & Blanco, 2014).

![Figure 2: SAE International J3016 standard of automation levels](https://dryve.com/glossary/what-are-the-sae-automation-levels).

### 2.3 Theoretical models of automation

Frameworks and models of automation are often produced for design and engineering purposes and focus on machine factors. However theoretical models are increasingly considering the key variables that impact human behaviour interaction with automation. These models frequently include trust as a factor in the effective performance of an automation system. One recent example is that produced by Stowers, Oglesby, Sonesh, Leyva, Iwig, and Salas (2017). Designed to support engineers evaluating human-machine systems, their model considers inputs and processes which collectively influence safety and performance (See Figure 3). In this model, trust is considered to be
a precursor to the acceptance of automation, which combines with other human processes and states to influence automation use.

![Diagram of factors affecting safety and performance in human-machine systems](image1)

**Figure 3.** Framework of factors affecting safety and performance in human-machine systems (Stowers, Oglesby, Sonesh, Leyva, Iwig, & Salas, 2017).

![Diagram of human-autonomy system oversight model](image2)

**Figure 4.** Human-autonomy system oversight model (from Endsley, 2017).

Endsley’s (2017) human-autonomy systems oversight model provides a more granular model of the relationship between system autonomy characteristics and human cognitive functions (see Figure 4). It considers in more detail the precursors and successors of factors influencing
Automation interaction and performance, including the role of trust. In this model, trust is influenced predominantly by automation characteristics and situational demands upon the human operator. Trust in turn influences attention allocation, that is, the level of human monitoring of the automated system.

A number of similar models have been proposed (e.g., Sanchez, 2009). These models show fairly high convergence in the factors considered important, and they tend to differ only in the detail of factors considered or their placement in the overall model. Although these models provide a high-level view of the role of trust in automated system interaction, they do not consider the antecedents and effects of trust in detail. Theoretical models which focus specifically on these factors are considered later in this review.

2.4 Appropriate use of automation: use, misuse, disuse, abuse

Most autonomous systems still operate with humans as supervisory controllers who direct performance or collaborate as teammates (Endsley, 2017), and the most effective outcomes result from collaboration and coordination between humans and machines (Murphy & Shields, 2012). However, as autonomous systems become more complex, the ability of human operators to effectively interact becomes more challenging. Endsley (2017) defines this as an automation conundrum: as a system becomes more autonomous and its reliability and robustness increase, it becomes less likely that a human operator will be able to take over manual control when needed. Compounding this is automation brittleness, in which automation operates well for the range of operations for which is was designed but requires human intervention for situations not covered by its programming (Woods & Cook, 2006). Additionally, autonomous systems which incorporate sensor input produce information which is inherently uncertain or incomplete (Antifakos, Kern, Schiele, & Schwaninger, 2005). Bainbridge (1983) identified as an irony of automation the fact that, as automation becomes more complex, the role of the human operator often becomes more, rather than less, important.

In an influential paper Parasuraman and Riley (1997) integrated previous research on human interaction with automation and identified several issues that explain why use of automation can result in failure to deliver its potential benefits. Identifying disparities between expectations of automation designers and operator requirements, they categorised these disparities as misuse, disuse, and abuse.

Automation use refers to the human operator’s engagement of automation to perform tasks (often tasks they might otherwise perform manually) (Parasuraman & Riley, 1997). Appropriate use of automation can enhance performance and safety and, with care, automation systems can be designed and used in ways which maximise the capabilities and minimise the limitations of both human and automation (Vagia, Transeth, & Fjerdingen, 2016).

Disuse of automation refers to instances in which a human operator fails to utilise automation when it could enhance performance (Parasuraman & Riley, 1997). In particular, poorly performing automation or high rates of false alarms of an automation’s system warnings can undermine acceptance or trust in the automation, resulting in disuse.
Misuse refers to over-reliance on automation (Parasuraman & Riley, 1997). One type of misuse arises from a lack of monitoring (often referred to as complacency) by the human operator, resulting in the neglecting of automation failures or errors. This neglect may give rise to two types of human error, known collectively as operation bias (Mosier & Skitka, 1996). The first is omission error: when the human operator’s over-reliance on the automation results in a failure to notice an automation error if the automation does not alert them to it. The second is commission error: when the human operator follows recommendations given by the automation, despite the recommendation being wrong (Bahner, Elepfandt, & Manzey, 2008). These monitoring failures are more likely to occur when the task load of the human operator is high, and when the automation failures are not clearly indicated to the human. They are also more likely to occur when failure frequency is low, and less likely when failure frequency is high. This misuse may also be described as overuse, and it is frequently due to over-trust in the automation. Misuse also occurs when human operators use the automation in ways not anticipated by the designers (Lee, 2008).

Automation misuse and disuse describe two types of inappropriate reliance. Understanding how to reduce misuse and disuse is an important consideration in automation design and training (Lee & See, 2004). Misuse and disuse arise from a complex interaction of many factors such as workload and self-confidence, however it is trust which has emerged as a particularly important factor (Lee & See, 2004). Over-trust leads to misuse, and under-trust results in disuse (Lee & Moray, 1992; Muir, 1987). Facilitating appropriate levels of trust in automation can reduce the frequency of misuse and disuse, and therefore trust is key to improving productivity and safety in human-automation teams (Hoff & Bashir, 2015).

Automation abuse refers to instances in which automation is designed and implemented with a technology-centred focus, without regard for its effect on the human operator (Parasuraman & Riley, 1997). The replacement fallacy underlying automation abuse assumes failures can be avoided by automating the role of the operator. This ignores the role of the operator plays in maintaining an effective system and in dealing with unexpected events which the automation cannot accommodate (Woods & Dekker, 2000).

### 2.4.1 Human-machine teamwork

As automation expands in use, scope, and complexity, the role of the human operator has increasingly shifted from that of principal controller, to that of an active team-mate or partner who shares the control with the automation (Klein, Woods, Bradshaw, Hoffman, & Feltovich, 2004). Accordingly, research is increasingly seeking to examine human-automation team dynamics during task completion, including in multi-automation teams. Challenges being addressed include the ability (for either party) to attend to other team members, predict team member’s actions, understand the intentions of team mates, and to communicate their intentions to other team members. Automation which is unobservant and incomprehensible (which Woods refers to as “strong silent automation”) to a human team member becomes frustrating, ineffective, and potentially dangerous (Bradshaw, Hoffman, Johnson, & Woods, 2013). An emerging field of study is examining the ‘etiquette’ of human-automation collaborations, and how this affects performance (de Visser & Parasuraman, 2010).

---

1. Note that Dekker and Hollnagel (2004) caution against unconsidered use of terms such as ‘complacency’, cautioning against mistaking the label for the covert information processes that underlie behaviours.
3 Trust

Trust and its role in mediating relationships has been extensively examined in a wide range of fields, including psychology, sociology, human factors, philosophy, economics, and political science. Although the significance of trust in cooperative relationships is widely acknowledged (e.g., Rotter, 1967), there are many different conceptualisations and definitions of trust. Some commonalities exist among these: the majority of trust definitions across the various fields include a trustor to give trust, a trustee to receive trust, and a goal which is to be achieved through relevant behaviour or actions (e.g., Castelfranchi & Falcone, 2000; Mayer, Davis, & Schoorman, 1995). Most definitions additionally propose that trust can only exist in situations of uncertainty and vulnerability. Trust is necessary when there is an element of risk arising from the possibility that the trustee will fail to complete the task (Hardin, 2006). Trust also acts as a decision heuristic to facilitate choice when a complete understanding is impractical or not possible (Kramer, 1999; Lee & See, 2004).

Despite these shared aspects, inconsistencies exist between trust definitions. These have important theoretical and practical implications for the study of trust in automation, as they lead to an inability to validly compare, contrast and build upon previous research. One of the major inconsistencies between trust definitions is whether trust is considered to be an attitude, intention, or behaviour. For example, Mayer, Davis, and Schoorman’s (1995) widely-used definition views trust as an intention: “the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that party” (p. 712). Alternative definitions focus on rational choice behaviour to view trust as the observable choice behaviour which is a product of believing that the benefits of trusting will outweigh the costs (Adams et al., 2003). Still other definitions characterise trust as an attitude, for example, the “expectancy held by an individual that the word, promise, or written communication of another can be relied upon” (Rotter, 1967, p.651).

Lee and See (2004), in an influential integrated review of early research on trust and reliance on automation, utilised Fishbein and Ajzen’s (1975) Theory of Reasoned Action to reconcile conflicting definitions. The theory provides a framework outlining the relationship between attitude, intention, and behaviour, and has been used extensively as the basis of research on attitudes and behaviour in a wide range of domains. According to the framework, an affective evaluation of beliefs and experience determine attitudes, which in turn inform intentions. Behaviours result from these intentions. Lee and See (2004) propose that, within this model, trust affects behaviour as an attitude, rather than as a belief, intention or behaviour. Beliefs underlie trust, and different levels of trust may result in varying intentions and behaviours. They additionally note that, in relation to automation, defining trust in automation as an intention or a behaviour could result in confusing its effect with other factors which may also impact upon behaviour, such as workload, situation awareness, or time pressures.
This view is supported by Castelfranchi and Falcone (2000) who define trust, in both other people and in non-human agents, as a cognitive process: “a complex mental attitude of an agent $x$ towards another agent $y$ about the behaviour/action $\alpha$ relevant for the result (goal) $g$.” As the action is useful to $x$, and $x$ is relying on it, $x$ can be considered to be delegating the action or goal to $y$, that is, relying on $y$ to complete the action. Therefore, trust is the mental counterpart of delegation or reliance (Castelfranchi & Falcone, 2000).

Although cognitive attitude and behavioural action are easily distinguished, it is more difficult to separate the concepts of trust as an attitude and trust as an intention. This has implications in particular for the measurement of trust over time. Fishbein and Ajzen’s (1975) Theory of Reasoned Action proposes that intentions guide goal-directed behaviour, existing at an intermediate level of abstraction between abstract attitudes and concrete actions (Bagozzi, 1981). Thus, intentions are the immediate antecedents of behaviour, and may vary rapidly over time due to variations in outcome, or environmental or circumstantial factors in addition to variations in attitude. Attitudes, on the other hand, are built on beliefs, and hence will change more slowly due to ongoing cognitions and emotions regarding the outcomes of behaviours, among other influences. Therefore, defining trust as an intention may possibly lead to different measurements and results to those produced when trust is defined as an attitude. Whether this has practical importance has not been examined within the automation literature.

Lee and See’s (2004) comprehensive examination of trust and its effect on reliance on automation has formed the basis of much subsequent work examining trust in automation. However the defining and understanding of trust remains the subject of extensive ongoing debate in many fields of research, including automation (e.g., Costa, Fulmer, & Anderson, 2018; Eikeland & Saevi, 2017). As noted by Lewicki, Tomlinson, and Gillespie (2006), the vast literature on trust in different settings, and the varying distinctions made about types of trust and associated trust indicators, are so entwined it is difficult to disentangle them to discriminate between them coherently.
4 Trust in automation

4.1 The importance of examining trust in the use of automation

The increasing complexity of automation and autonomous systems has the consequence that it is often no longer possible for a human operator to know or understand all properties of the system or predict when faults will occur. Automation has the potential to increase performance and safety of task outcomes, however the human operator is dependent upon the automation functioning as required, creating a situation of vulnerability and uncertainty. Irrespective of the robustness of automation design, it is likely that it will in some instances fall short of expectations. Thus, it is not possible for an operator to simply use the automated system when it works, and desist when it is faulty (Muir, 1994). Sheridan (1975; Sheridan & Ferrell, 1974; Sheridan & Hennessy, 1984) has long argued that trust is key in mediating the relationship between humans and automation, functioning in a similar manner to trust between humans. When trust in automation exceeds the human operator’s self-assessed ability to perform a task, the automation is likely to be used, but if self-confidence exceeds trust, then the human operator is unlikely to use the automation (Lee & Moray, 1994).

4.2 Trust in people generalised to trust in automation

A number of studies have demonstrated that human responses to technology closely mirror responses to other humans. For example, studies have demonstrated that people use socially acceptable behavioural rules such as politeness when interacting with technology (Reeves & Nass, 1996), and readily form team relationships with computers (Nass, Fogg, & Moon, 1996). Nass, Moon, and Carney (1999) argue that the similarity of cues in the human-computer interaction to cues normally associated with human-human interactions elicit automatic and unconscious processes. When presented with human-computer cues such as the use of language, collaborative interactions, and the filling of roles traditionally filled by humans, their similarity to human-human cues lead people to automatically utilise established schema associated with interpersonal interactions.

Such parallels have been shown to extend to trust (Muir, 1994). Research has demonstrated a strong qualitative similarity between human-human trust and human-automation trust, in particular, a correspondence between the ongoing dynamics of trust during task completion in a complex environment (Lewandowsky, Mundy, & Tan, 2000). Trust in both types of relationship develops typically as a function of trustee characteristics and behaviours (Madhavan & Weigmann, 2004). Madhavan and Wiegmann (2007) undertook an integrative review of literature to specifically examine similarities and differences between human-human trust and human-automation trust in team structures. They note that automation such as decision aids are often designed specifically to
behave in a manner similar to humans through language structures and content, resulting in similar trust attitudes in the human operator.

Differences do seem to exist between human-human trust and human-automation trust, however. Trust in automation does not yet display the symmetry of mutual awareness in the behaviours and intents of the other (Lee & See, 2004). In response to this, some researchers have suggested that trust in automation actually embodies trust in the designers behind the automation (e.g., Parasuraman & Riley, 1997). Another difference exists in the effects on trust in automation due to design and interface characteristics such as menus and graphics, which have no equivalent in human-human trust dynamics (Corritore, Kracher, & Wiedenbeck, 2003). The inclusion of anthropomorphic features in autonomous systems may circumvent this difference, however the inclusion of such features has the potential to attribute intentionality where none exists (Lee & See, 2004; Waytz, Cacioppo, & Epley, 2010).

4.3 Defining trust in automation

The use of precise definitions of trust varies widely in automation trust literature, with many empirical studies failing to define trust at all. Lee and See (2004) propose a simple definition of trust which is consistent with their evaluation of trust as an attitude: “the attitude that an agent will help achieve an individual’s goals in a situation characterised by uncertainty and vulnerability” (p. 54). Since its publication, this definition of trust is by far the most widely used in empirical studies of trust in automation. The definition proposed by Mayer, Davis, and Schoorman (1995; see above) is also used in relation to trust in automation. However, generally across the trust in automation literature, the lack of consistency in trust definitions, and the imprecision in measuring trust that can result from this, remain an ongoing issue with consequences for comparisons across studies (McKnight & Chervany, 2001).

4.3.1 Links with similar concepts

Trust can be difficult to disentangle from related concepts, which has led to difficulty in interpreting results across studies. Despite conceptual differences discussed within trust literature in other fields, the terms below are frequently used interchangeably with trust within automation literature, and few studies explicitly define differences between them.

4.3.1.1 Reliance

Lee and See (2004) argue that reliance is a behavioural result of the attitude of trust. In trust-in-automation studies, reliance on automation is frequently employed as a proxy measure of trust. Lee and See (2004) note that this can have the potential to attribute reliance behaviours to trust when they may be due to other factors such as workload or situation awareness. Gaudiello, Zibetti, Lefort, Chetouani, and Ivaldi (2016) also argue for differences between trust and reliance. In relation to robots, they note that trusting a robot implies a belief that the robot is able to correctly perform a task while not necessarily depending on the robot to accomplish that task. For reliance, the operator does depend on the robot to accomplish the task.
4.3.1.2 Risk

While trust (an attitude) towards an agent may make an individual willing (an intention) to be vulnerable, risk is intrinsic to the actual manifestation (a behaviour) of that intention (Mayer, Davis, & Schoorman, 1995). Thus, risk and reliance are closely associated. However, trust is not a necessary precursor of risk; risk-taking behaviour can occur in the absence of trust.

4.3.1.3 Cooperation

Trust can often result in cooperative behaviour. However, as cooperation does not necessarily involve an element of vulnerability, trust is not a necessary condition for cooperation to occur. (Mayer, Davis, & Schoorman (1995).

4.3.1.4 Confidence

Mayer, Davis, and Schoorman (1995) note that, in literature on trust, the relationship between confidence and trust is imprecise, with each term often being used when defining the other. Luhmann (1988) argues that trust differs from confidence in the attributions and perceptions involved. An attitude of confidence means that alternative options are not considered. In contrast, an attitude of trust involves choosing one action in preference to others despite the possibility of being let down by the actions of another. Siegrist, Earle, and Gutscher (2003) support this distinction, defining confidence as a ‘belief, based on experience or evidence, that certain future events will occur as expected’ (p706). While trust involves risk and has a basis in relationships, confidence does not involve risk, and is based on familiarity, prior experience, and base rates. The authors state that the objects of trust are person-like entities, while confidence can be bestowed upon virtually anything. They note that trust and confidence may interact, but that little is known about this process, although in later work they define confidence as ‘trust in competence’, determined solely by previous experience of the actions of the trustee (Siegrist, Gutscher, & Earle, 2005). Adams (2005) notes that while trust utilises cognitions, emotions, and motivations to make discrete or holistic judgments, confidence is based only upon cognitions to make discrete judgments only. Hoffman, Lee, Woods, Shadbolt, Miller, and Bradshaw (2009) additionally point out that it is possible to have confidence that automation will not work, a state they term ‘antitrust’.

4.3.1.5 Predictability

As with confidence, the relationship between predictability and trust is ambiguous, with many papers equating the two (Mayer, Davis, & Schoorman, 1995). While prediction and trust are both means of reducing uncertainty, trust must go beyond predictability to be meaningful (Deutsch, 1958). An agent may persistently produce negative outcomes, and hence be predictable. It is unlikely, however, that an individual will be willing to be vulnerable to, and so place trust in, this agent. Predictability may instead be best envisaged as influencing co-operation rather than trust (Mayer, Davis, & Schoorman, 1995).

4.3.1.6 Trustworthiness

Trustworthiness involves only attributes of the trustee, rather than the individual bestowing trust (Mayer, Davis, & Schoorman, 1995). Trust formation (discussed below) rests largely upon the perceived trustworthiness of the automation.
4.3.1.7 Trust, distrust, mistrust

Mistrust and distrust are often used interchangeably, with distrust being the term more commonly used. There is disagreement about the differentiation between trust and distrust. Through factor analysis Jian, Bisantz, and Drury (2000) concluded that trust and distrust can be considered opposite ends of the same construct rather than conceptually different. This view is supported by a number of other researchers (e.g., Rotter, 1971; Schoorman, Mayer, & Davis, 2007). However, Slovic’s (1993) asymmetry principle asserts that they are not the inverse of each other, as trust and distrust are learned in different ways, and function differently. Slovic proposes that this is the reason people subjectively view trust-destroying events as carrying much greater weight than trust-building events, and why trust can be instantly destroyed while distrust, once initiated, tends to reinforce itself. Lewicki, McAllister and Bies (1998) also view trust and distrust as distinct concepts, with trust focusing on the possibility of a desirable outcome and distrust focusing on the possibility of an undesirable outcome. Muir (1987) defines mistrust as the error of either distrusting competent automation, or of trusting incompetent automation (this generally now referred to as over-trust and under-trust, and is discussed in greater detail further below). Within the trust in automation literature, research more commonly focuses on different levels of trust and how these levels change, either over time or due to different factors, rather than on distrust. Changes in level of trust are commonly referred to as losses/decline or gains/growth of trust.

4.4 Theoretical basis of trust formation

Rempel, Holmes, and Zanna (1985) proposed that trust is a dynamic attitude which follows a particular sequence of dimension to form gradually over time. They identified three dimensions which influence an individual’s acceptance of a trustee to form the basis of trust: predictability, dependability, and faith. This model was subsequently applied to trust in automation by Muir (1987; 1994), Lee and Moray (1992), and Lee and See (2004). A very similar model, developed for considering organisational trust, refers to the three bases of trust as ability, benevolence, and integrity (Mayer, Davis, & Schoorman, 1995). When theoretical foundations of trust formation are considered in automation literature, it is most commonly through one of these very similar frameworks.

Predictability forms the initial basis of trust, and it depends predominantly upon the stability of performance over time (Rempel et al, 1985). If a trustee’s behaviour is assessed as both consistent and desirable, the perception of predictability is enhanced. Assessment of predictability depends upon the actual predictability of the trustee’s behaviour, the trustor’s ability to accurately assess the predictability of the trustee, and the stability of the environment in which trust is occurring. Automation which performs consistently as expected and is transparent (easily observed and understood) will produce a positive assessment of predictability by the operator and lead to initial trust (Muir, 1994). Experience either of manually undertaking the task being performed or with similar automated systems may also enable an experienced operator to more accurately assess the automation’s predictability than novice operators. In considering effects of the environment in which the automation operates, operators must distinguish between unpredictability resulting from an unstable environment from inherent unpredictability unrelated to environmental changes. The
former should not negatively impact trust, while the latter should (Muir, 1994). Lee and Moray (1992) refer to this stage as ‘performance’ and include automation reliability and ability, in addition to predictability, as characteristics influencing this stage of trust development. Performance evaluations are made in relation to the specific goals of the operator, demonstrating the task-dependent nature of trust (Lee & See, 2004).

*Dependability* forms the basis of trust after a period of time and experience, as attention refocuses from assessment of specific behaviours to an assessment of dispositional characteristics of the trustee, most particularly on their dependability, or the degree to which they can be relied upon (Rempel et al., 1985). An attribution of dependability is formed from an accrual of positively predictable behavioural evidence, with particular focus on events involving risk or vulnerability. To attribute dependability to automation, a considerable amount of experience with the automation, including in uncertain scenarios, is required (Muir, 1994). Once an operator determines that the automation is dependable, the close monitoring required in the earlier stage is no longer required for trust to be maintained. In Lee and Moray’s (1992) version this stage, denoted ‘process’, is related to the extent to which the automation’s design is appropriate for achieving the operator’s goals. While performance describes *what* the automation does, this stage refers to *how* the automation operates, and the focus shifts from specific behaviours towards qualities and characteristics attributed to the automation. The operator at this stage will tend to trust the automation if its processes are comprehensible and appear able to achieve the operator’s goals, within the operating environment (Lee & See, 2004).

*Faith* is the final stage in this model of trust formation (Rempel et al., 1985). Faith is based on beliefs about the future behaviour of the trustee. Past predictability and dependability are used as a basis for belief that the trustee will behave in the future as they have in the past, with additional consideration given to the trustee’s perceived motives. In automation, many automated processes are too complex for the operator to have a complete understanding of them, and will potentially require unanticipated interaction (Muir, 1994). To effectively control a system under uncertainty requires a leap of faith beyond the behaviour evidence accrued previously. Lee and Moray’s (1992) version refers to this stage as ‘purpose’, and they consider it to be the extent to which the automation is being used in line with the designer’s intent. This stage, then, refers to *why* the automation was created, and this stage of trust attribution frequently depends on whether the operator understands the designer’s intent for the automation. If this is the case, the operator will tend to trust the automation to successfully accomplish the goals for which it was designed (Lee & See, 2004). Thus trust is both an *outcome* of automation characteristics such as reliability or robustness and a *cause* of a human operator’s behaviour when using automation (Sheridan & Parasuraman, 2005).

A few empirical studies have examined this trust formation process in relation to technology and automation. For example, a field study examining trust in an online system revealed three phases of trust, common to both new and experienced users of such systems (Söllner & Pavlou, 2016). During the first three weeks of use, users confirmed whether their initial trust evaluation was correct, and adjusted their level of trust based on their experience of the system. After this phase, trust in the system grew steadily. After an additional six weeks, this growth plateaued and remained at a stable level.
There is considerable agreement that the formation and tuning of trust is a result of both cognitive and affective processes (Lewis & Weigert, 1985; McAllister, 1995). When adequate cognitive resources are available individuals often use deliberate, analytic thought processes to decide whether or not to trust. When cognitive resources are limited, mental heuristics or emotions are more likely to influence trusting behaviour (Hoff & Bashir, 2015). Cognitive-based trust processes are based on knowledge of the trustee’s competence, reliability, and dependability (Luhmann, 1979), while affective-based trust processes are based on emotional investments, concern for others, and the belief of reciprocity of sentiment (Lewis & Weigert, 1985). Typically, both cognition and emotion play a role in an act of choosing to trust (Zajonc, 1980). Many authors consider trust in automation to be largely formed from cognitive processes based upon the automation’s responses (e.g., Muir, 1994).

However, Lee and See (2004) propose that trust is predominantly an affective response, but that this emotive response is influenced by cognitive processes, which they divide into analytic and analogical categories. Analytic processes involve rational weighting of evidence through reasoning. Analogic methods use recommendations of others and presumptions based upon category memberships to infer trust levels.

Few studies thus far have examined the possible effects of using these different mental processes on trust in automation. Parasuraman and Miller (2004) examined the role of social etiquette in human-automation trust. Manipulating the communication style of a flight simulator when delivering information text messages to the human operator, they adjusted the propensity of the automation to interrupt or appear impatient. This was intended to alter the social etiquette acceptability of the automation, thus affecting the human operator’s trust through analogic or affective mental processes. Although a strong effect on trust was observed, with positive etiquette features increasing trust and negative etiquette features decreasing trust, it was not possible to definitively ascertain the mental processes involved, and hence it was simply deduced that affective processes were most likely involved.

Many factors influence the development of trust and changes in trust over time. These factors and their effect on trust, both as it is measured at one point in time and as it changes over time, are examined in a separate section below.

### 4.5 Appropriate calibration of trust in automation

For a human-automation team to successfully accomplish its goals, the human operator must rely appropriately on the automation. Over-reliance may result in misuse, with the operator depending upon the automation to perform outside its parameters or when faulty. Under-reliance may lead to disuse, resulting in excessive operator workload and diminished system performance (Lee & Moray, 1992). Appropriate reliance on automation depends to a large extent upon how well trust in the automation matches the automation’s actual performance (Lee & See, 2004). This correspondence between a human operator’s trust in automation and the actual capabilities of the automation is referred to as **calibration** (Muir, 1987). Appropriate calibration occurs when the level of trust matches the capabilities of the automation (see Figure 5). It reflects the human operator’s accurate understanding of the level of imperfection inherent in the automation, and should result, where
necessary, in corresponding adjustment in the interactions with the automation to maintain optimum performance and safety (Wickens, Gempler, & Morphew, 2000). Xu, Wickens, and Rantanen (2007) have demonstrated that, once a human operator has reached an appropriate calibration of trust, he or she can still effectively complete tasks even with imperfect automation.

![Figure 5. Representation of calibration of trust in automation for appropriate reliance (adapted from Lee & See, 2004).](image)

**Overtrust** is poorly calibrated trust in which the level of trust overestimates the capabilities of the automation. Overtrust can result in misuse of automation. **Undertrust** (initially termed distrust by Lee & See, 2004, but later undertrust in Lee, 2008) is poorly calibrated trust in which the level of trust underestimates the automation’s true capabilities. Undertrust can lead to automation disuse. The results of overtrusting or undertrusting automation can be catastrophic with life-threatening consequences, as the many accounts included in automation trust literature testify. A great deal of research on automation trust focuses either on factors which affect trust calibration, or on the antecedents and effects of overtrust or undertrust. These are discussed in a separate section below.

Less extensively utilised in automation trust literature, Lee and See (2004) suggest additional gauges for examining appropriate levels of trust. **Resolution** refers to the extent to which a judgement of trust discriminates between different automation capabilities, that is, differences in how well the automation performs. Low resolution occurs when large changes in automation capability result in only small changes in trust. **Specificity** describes the extent to which trust is dependent upon a particular component or aspect of the automation, and it is divided into functional specificity (associated with the functions and modes of automation operation), and temporal specificity (changes in trust over time or as a function of the situation). High functional specificity relates to trust in specific tasks or functions of the automation, while low functional specificity reflects trust in the capabilities of the overall system. High temporal specificity results from trust reflecting moment-to-moment variations in automation capability. Trust reflecting long-term changes in automation capability is referred to as low temporal specificity.

Recent work by Merritt, Lee, Unnerstall, and Huber (2015) attempts to link resolution and specificity with appropriate trust calibration by dividing calibration into three categories. **Perceptual accuracy** refers to the extent to which the user’s perception of performance matches actual automation performance, while perceptual sensitivity and trust sensitivity refer to the user’s
adjustment over time of perceived reliability and trust respectively. They conclude that trust calibration warrants further investigation, as moderators of appropriate trust calibration may exist.
5 Models of trust in automation

A number of theoretical and quantitative models of trust in automation have been developed over the past 25 years.

5.1 Early models

5.1.1 Muir (1994), Muir and Moray (1996)

Muir (1987) argued that automation systems were being developed with little consideration for their actual need. She argued that automation development was being driven by technology rather than objectives or design principles, resulting in systems in which the emphasis was placed on performance of the automation at the expense of the human operator. To address this, she developed a framework for the study of trust in automation based on work on interpersonal trust by Barber (1983) and Rempel, Holmes, and Zanna (1985; Muir, 1994). This framework closely matches the theoretical basis of trust formation outlined above (predictability, dependability, faith). Muir additionally proposed that trust formation depended upon three human expectations: persistence (in the natural and moral social orders), technical competence (automation performance demonstration over time), and fiduciary responsibility (that the other agent is morally obligated to prioritise the other’s interest before their own). Muir noted that this framework was theoretical, and that a simpler model may still provide a good fit. Subsequent trust frameworks have borne out this view.

From this trust framework, Muir (1994) developed a qualitative model of the relationship between automation, operator’s trust, and automation behaviour predictions in order to examine operator trust calibration (see Figure 6). The model was intended to provide a theoretical framework for planning, interpreting, and integrating future research in trust in automation. The model is significant in being the first major contribution to the field, and many of the model’s theoretical links have been empirically demonstrated, however it has since been superseded by models which better encapsulate trust in more complex automated systems (Adams et al., 2003).
5.1.2 Other early models

Early models of trust in automation built on the ongoing work in the field to gradually create more empirically justified models. The following models present a flavour of this gradual development.

Lee and Moray (1992) developed a time series model to represent the dynamics of trust during fault management. The model represents the first attempt to model trust in a system dynamically to consider how trust is affected by system faults in addition to performance, however it is limited in the conclusions that can be drawn from it (Adams et al., 2003).

Cohen, Parasuraman, Serfaty, and Andes (1998) developed a computational model of trust in decision aids. Their model was created with the aim of developing a probabilistic theory which could consider the contextual and temporal elements of trust (Adams et al., 2003), however the model was not widely adopted, perhaps because it was formulated specifically for automation use within armed forces missions and training.
Madsen and Gregor’s (2000) model of human-computer trust was created to assist in the development of a subjective scale for measuring human-computer trust. Based solely on attributes of the human operator, it divides trust into cognitive components (perceived understandability, technical competence, reliability) and affective components (personal attachment, faith) and defines trust as the user’s confidence in the system and their willingness to act on the system’s advice. The model is notable for being one of the first to distinguish explicitly between cognitive and affective aspects of trust, but it has not been further developed or adopted.

Seong and Bisantz (2000) developed a model of human trust in automation based on the Lens model, which represents information transformation from stimulus to response. It theorises that the human operator bases trust on observable characteristics of the automation, such as its reliability and dependability. Subsequent use of the automation is theorised to reflect the level of trust. This is problematic, as subsequent research has clearly shown an effect on use of automation of factors other than trust (Adams et al., 2003).

Kelly, Boardman, Goillau, and Jeannot (2003) developed a model of trust in automation which outlined factors that impact upon trust, and the relationship between them. The model proposes three main sources of trust: understanding, competence of the automation, and self-confidence. Understanding is built on predictability, explication of intention and familiarity. Automation competence is assessed through its dependability, reliability, usefulness, and robustness. Self-confidence is developed through faith, reputation, skills and training, and personal experience. Strengths of the model include the distinction it makes between the user’s perception of automation performance and actual performance, and the fact that it includes training as an influencing factor on trust in the automation, albeit only indirectly through self-confidence (Adams et al., 2003).

Adams and colleagues (2003) created a model of trust in automation for use within the Canadian military. Their model detailed both the development of trust and the factors affecting trust in automation (see Figure 7). The model is notable for its consideration of factors influencing the trustor, the automation, and the context. This distinction between types of factors influencing trust has been used in a number of subsequent models. The potential conflict between factors evaluated by the trustor, and contextual factors imposed upon the trustor is also a novel distinction addressed by this model.
Nickerson and Reilly (2004) created a model for investigating the effects of machine autonomy on human behaviour which focused on initial trust, and considered cognitive, affective, and dispositional factors including, unusually, the role of stress. Unfortunately for the authors, this model was published in the same year as one of the most influential papers in the trust in automation field, and this model has received little further consideration.

5.2 Recent and current models

5.2.1 Lee and See (2004)

The review of trust in automation by Lee and See (2004) greatly influenced subsequent trust in automation research. The model they developed as a consequence of this review considers the dynamics of trust in automation and its effect on reliance upon automation. In a similar manner to the popular Technology Acceptance Model (Davis, 1985), the model was based on the Theory of Reasoned Action/Theory of Planned Behaviour (Ajzen, 1991; Fishbein and Ajzen, 1975), and was intended to complement an existing framework that models factors affecting automation use (Dzindolet, Pierce, Beck, & Hall, 2002). Lee and See’s closed-loop model proposes a dynamic interaction among the context, the human operator, the automation, and the automation’s interface (see Figure 8).
Figure 8. A conceptual model of the dynamic process that governs trust and its effect on reliance (from Lee & See, 2004).

The model details a closed-loop process in which interaction with the automation influences trust, and trust influences interaction with the automation (Lee & See, 2004). Contextual factors in the top box influence beliefs (in the left box of the middle row) from which trust attitudes evolve. From a combination of these attitudes and situational factors, an intention to rely (or not) on the automation is formed. If the intention is to rely and contextual and automation performance allows, this intention is transformed into behavioural reliance and the automation is used and trusted. Following use, information about the automation’s performance is assessed through the automation’s display and feeds back into beliefs about the automation. Therefore, changes in level of trust depend on the combination of operator characteristics, automation characteristics and the context, rather than any one element alone.

Aspects of the model have proved influential in subsequent work, in particular, the dynamic interaction between person, automation, and contextual factors to influence trust over time. However, the model as a whole does not appear to be widely used as a basis for empirical research. Distinctions between belief, attitude, and intention are difficult to ascertain in an experimental setting, and much subsequent work in the area simply appears to collapse these three aspects into the simple term ‘trust’. As previously mentioned, there may be distinctions to be drawn between intentions to trust, which may change rapidly, and attitudes of trust, which may alter more slowly. However, it is unclear if this is of practical importance.

Lee has subsequently described amendments he would make to the model (Atkinson, Hancock, Hoffman, Lee, Rovira, Stokes, & Wagner, 2012). He proposes that, while interaction might
previously have been described in terms of reliance and compliance, it may now be better
described in terms of cooperation and may be more likely to include aspects of social exchange
relationships. Inclusion in the model of the effects of personal and developmental history, and of
other relevant stakeholders has also been suggested (Ho, Sadler, Hoffmann, Lyons, & Johnson,
2017).

Perhaps to appeal more to a technical research community, Ghazizadeh, Lee and Boyle (2012)
produced a conceptually similar model which was more explicitly based on the Technology
Acceptance Model (Davis, 1985) (see Figure 9). The Technology Acceptance Model (Davis, 1985)
has been criticised for its limited explanatory and predictive power, and lack of practical value (Chuttur,
2009). The Automation Acceptance Model, although it has been subsequently utilised in a number
of research papers, appears to suffer from the same limitations.

Theoretical models developed subsequently to Lee and See’s (2004) model have largely followed
the format of separating the factors influencing trust in automation into the categories of human
operator, automation, and environment/situation/context. The models differ in whether they
explicitly model the dynamic nature of trust over time, or whether it is simply implied.

5.2.2 Hancock, Billings, Schaefer, Chen, de Visser, and Parasuraman (2011)

Hancock et al. (2014) developed a model which divided factors of trust development in human-
robot interaction into human-related, robot-related, and environmental categories (see Figure 10). The
model was intended to inform a subsequently performed meta-analysis, but has been more
widely used.

The model was later assessed by some of its authors to be limited in its scope, and so was adapted to
consider factors associated with trust in automation more generally (Schaefer, Billings, Szalma,
Adams, Sanders, Chen, & Hancock, 2014). This was intended to incorporate empirical evidence of
the wider automation literature, but the aim of the model was to further increase understanding
specifically in human-robot trust (see Figure 11).
**Figure 10.** Factors of trust development in human-robot interaction (from Hancock et al., 2011).

<table>
<thead>
<tr>
<th>Human-Related</th>
<th>Robot-Related</th>
<th>Environmental</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ABILITY-BASED</strong></td>
<td><strong>PERFORMANCE-BASED</strong></td>
<td><strong>TEAM COLLABORATION</strong></td>
</tr>
<tr>
<td>Attentional Capacity/Engagement</td>
<td>Behavior</td>
<td>In-group Membership</td>
</tr>
<tr>
<td>Expertise (Amount of Training)</td>
<td>Dependability</td>
<td>Culture</td>
</tr>
<tr>
<td>Competency</td>
<td>Reliability of Robot</td>
<td>Communication</td>
</tr>
<tr>
<td>Operator Workload</td>
<td>Predictability</td>
<td>Shared Mental Models</td>
</tr>
<tr>
<td>Prior Experiences</td>
<td>Level of Automation</td>
<td></td>
</tr>
<tr>
<td>Situation Awareness</td>
<td>Failure Rates</td>
<td></td>
</tr>
<tr>
<td><strong>ATTRIBUTE-BASED</strong></td>
<td><strong>TRANSPARENCY</strong></td>
<td><strong>TASKING</strong></td>
</tr>
<tr>
<td>Transparency</td>
<td></td>
<td>Task Type</td>
</tr>
<tr>
<td><strong>ATTRIBUTES</strong></td>
<td></td>
<td>Task Complexity</td>
</tr>
<tr>
<td>Demographics</td>
<td></td>
<td>Multi-tasking Requirement</td>
</tr>
<tr>
<td>Personality Traits</td>
<td></td>
<td>Physical Environment</td>
</tr>
<tr>
<td>Attitudes toward Robots</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comfort with Robot</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-confidence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Propensity to Trust</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 11.** Theoretical model of human-automation trust (from Schaefer et al., 2014).

<table>
<thead>
<tr>
<th>Human Operator</th>
<th>Automation</th>
<th>Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Human Traits</strong></td>
<td><strong>Features</strong></td>
<td><strong>Task-Related</strong></td>
</tr>
<tr>
<td>Age</td>
<td>Degrees of automation</td>
<td>++ Proximity</td>
</tr>
<tr>
<td>Gender</td>
<td>Appearance</td>
<td>+ Risk</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>Mode of communication</td>
<td></td>
</tr>
<tr>
<td>Personality</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Human States</strong></td>
<td><strong>Capability</strong></td>
<td></td>
</tr>
<tr>
<td>Fatigue</td>
<td>Errors in automation</td>
<td></td>
</tr>
<tr>
<td>Stress</td>
<td>Automation behavior(s)</td>
<td></td>
</tr>
<tr>
<td>Attentional control</td>
<td>Quality/Accuracy</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cueing/Feedback/Alarms</td>
<td></td>
</tr>
<tr>
<td><strong>Cognitive Factors</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Understanding the automation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ability to use automation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expectancy of automation</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Emotive Factors</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attitudes toward automation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Confidence in automation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Satisfaction with automation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comfort with automation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: * represents experimental and + represents correlation findings.
5.2.3 Hoff and Bashir (2015)

Hoff and Bashir (2015) produced perhaps the most comprehensive recent model to integrate empirical evidence on factors that influence trust in automated systems. The model is intended to be applicable to a range of automated systems and situations. It organises the factors influencing trust into the commonly used three categories of human operator, environment, and automated system. Relating these to layers of trust (Marsh & Dibben, 2005), they ultimately divide the factors into dispositional trust, situational trust, and learned trust.

Dispositional trust includes individual differences and long-term tendencies of a human operator (see Figure 12; Hoff & Bashir, 2015). Together, they form an individual’s overall tendency to trust automation, irrespective of the specific context or system. This form of trust is considered to be relatively stable over time.

Figure 12. Factors that influence dispositional trust (from Hoff & Bashir, 2015).

Situational trust refers to the collection of internal and external factors related to the particular conditions under which trust occurs (see Figure 13; Hoff & Bashir, 2015). These factors both directly affect trust in automation, and also partially determine the degree to which trust influences behaviour towards the automation. Internal variability factors differ from dispositional factors by being more transitory in nature.

Figure 13. Factors that influence situational trust (from Hoff & Bashir, 2015).

Learned trust refers to a human operator’s evaluations of a system based on past knowledge and experience, and current interaction with the automation (see Figure 14; Hoff & Bashir, 2015). It is influenced by the level of performance of the automation system. Design features can also influence learned trust, but they do so indirectly by affecting perceptions of system performance.
Figure 14. Factors that influence learned trust. The dotted arrows represent factors that can change within the course of a single interaction (from Hoff & Bashir, 2015).

The full model combines dispositional, situational, and learned trust (see Figure 15; Hoff & Bashir, 2015). In this version, factors which contribute to dispositional, situational, and initial learned trust determine initial reliance on the automation, while dynamic learned trust is represents subsequent variations in trust levels as automation interaction continues.

Figure 15. Full model of factors that influence trust in automation. The dotted arrows represent factors that can change within the course of a single interaction (from Hoff & Bashir, 2015). Additional factors not related to trust have also been included in the full model. This is to account for the fact that trust contributes, but does not completely determine, reliance on automation (Lee &
See, 2004). Additional factors, such as alternative job completion strategies, time constraints, or the operator’s situational awareness or physical well-being can also potentially influence reliance, irrespective of trust levels. Since publication the model has been used in numerous subsequent studies to support consideration of the impact of various factors on dynamic trust changes in automation.

5.2.4 Bindewald, Rusnock, and Miller (2018)

Bindewald, Rusnock, and Miller (2018) produced a model of trust behaviour in human-machine teams (see Figure 16). The model is intended to describe trust behaviour which goes beyond compliance and reliance to include situations in which the machine has authority to act on behalf of the human-machine team. The purpose of the model is to enable the development of trust measures through the measurement of behaviours which can be distinguished from those caused by other factors such as previous experience, mental capacity, and workload.

![Figure 16](image)

Figure 16. A causal loop diagram illustrating effects upon trust and reliance (from Bindewald, Rusnock, & Miller, 2018). (Note KSA = knowledge, skills, and abilities).

5.3 Models of specific facets of trust in automation

Madhavan and Wiegmann (2007) created a more unusual model which is intended to compare a human operator’s trust development in decision support systems to the operator’s trust development in another human (see Figure 17). The aim of the model is to improve the design of decision support systems by incorporating factors which elicit operator responses to the automation which more closely match human-human responses. In practice, this may prove difficult to achieve until technical advances in autonomous systems enable more nuanced interactions between human and automation. This model may, however, be beneficial in supporting emerging human-automation teams.
Figure 17. Model of sequential trust development in human vs automated decision support systems (DSS) (from Madhavan & Weigman, 2007).

Figure 18. Model of factors of trust development in automation. Performance expectancy, process transparency, and purpose influence are the constructs (solid lines); individual differences, cultural-technological contexts, and cultural differences are the moderators (dotted arrows) (from Chien, et al., 2014).
Chien, Lewis, Semnani-Azad, and Sycara (2014) created a model of cultural factors influencing trust in automation to support their development of a psychometric subjective measure of trust in automation (see Figure 18). The model maps cultural characteristics and individual differences onto the three bases of trust formation (see section: Theoretical Basis of Trust Formation). They use the nomenclature adopted by Lee and See (2004) to demonstrate how cultural and individual factors map onto the trust development stages of performance, process, and purpose. The model highlights the importance of considering the impact of cultural factors on trust levels, as substantial research has shown that trust (in general) varies across cultures and countries (e.g., Naef, Fehr, Fischbacher, Schupp, & Wagner, 2008), although more research is needed to specifically examine cultural effects on trust in automation (Hoff & Bashir, 2015).

Hoffmann and Söllner (2014) developed a model of trust formation. The model considers the antecedents of the three dimensions underlying the formation of trust: performance, process and purpose (Lee & See, 2004; see Figure 19). The model is intended to enable inclusion of insights from behavioural trust theory in the design of automation elements.

**Figure 19.** Model of the formation of trust in automated systems (from Hoffmann and Söllner, 2014).

### 5.4 Models of specific types of automation

Skarlatidou, Cheng, and Haklay (2013) proposed a simple model of trust in online trust components. The model separates characteristics of the trustor and online system, and additionally includes the influences of affective and cognitive trust and trust preconditions. Although it has been utilised in subsequent research, the model does not appear to offer the design or predictive ability of more comprehensive models.

Ekman, Johansson, and Sochor (2018) proposed a highly detailed framework for trust in automated vehicle systems (see Figure 20). The top section shows the ‘life cycle of trust’, which is a visualisation of the trust formation process. The middle section of the model details driving events. The bottom section shows trust-affecting factors which are associated with each driving event.
Peterson, Robert, and Yang (2017) have proposed research that aims to begin the process of building models of mutual trust between a human driver and a semi-autonomous vehicle. While still in very early stages of the process, models such as these have the potential to more effectively describe and predict interactions among human-automation teams of the future.
6 Factors influencing trust in automation

For trust to be appropriately calibrated, the operator should understand the conditions under which optimal system performance occurs (McBride & Morgan, 2010). Factors influencing trust in automation are commonly divided into three categories: characteristics of the operator, characteristics of the automation, and environmental influences. Evidence from a number of studies indicates that automation characteristics exert the largest effect on trust levels. For example, a recent meta-analysis looking specifically at robots found that robot attributes and performance were the largest contributors to the development of trust, with environmental factors having moderate effect, and human factors having little effect (Hancock, Billings, Schaefer, Chen, de Visser, & Parasuraman, 2011). This, however, must be considered in light of the fact that automation characteristics are also by far the most studied factors in relation to trust, with situational characteristics the least examined. Schaefer, Chen, Szalma, and Hancock (2017) propose that, as it becomes more possible to calibrate automated systems to the requirements of individual operators, additional research may demonstrate that operator characteristics may in fact exert some of the most significant influences. The manner in which trust is bestowed upon automation in real-world scenarios, however, is most likely to stem from a complex combination of person, automation and environment factors difficult to replicate in laboratory settings.

6.1 Operator factors

6.1.1 Culture

Culture is known to have significant effects on interpersonal trust (e.g., Naef, Fehr, Fischbacher, Schupp, and Wagner, 2008), although few studies have yet demonstrated specific cultural effects of trust in automation. Rau, Li and Li (2009), in an examination of the effect of culture on ability to accept recommendations from robots, reported that Chinese participants reported the robots as more trustworthy than German participants. Mexicans have been shown to be more likely to trust automated decision aids compared to Americans (Huerta, Glandon, & Petrides, 2012). Chien, Lewis, Sycara, Liu, and Kumru (2016) studied the relation between trust in automation, culture, and personality traits in US, Turkish, and Taiwanese participants, however results showed no universal patterns across the three cultures in their propensity to trust automation.

6.1.2 Age

Age-based differences in trust may result from some combination of cognitive changes or cohort effects (Steinke, Fritsch, & Silbermann, 2012). Some evidence suggests that age may influence what types of automated systems are trusted. Older adults have been shown to be more likely to trust driver warning systems (Donmez, Boyle, Lee, & McGhee, 2006) and medication management decision aids (Ho, Wheatley, & Scialfa, 2005) than younger adults, although no differences were
seen between the two groups in trust adjustments after automation errors. Adding a picture of a doctor to a diabetes management application’s interface resulted in younger participants placing higher levels of trust in the system’s advice, but no effect was observed for older participants (Pak, Fink, Price, Bass, & Sturre, 2012). Although results suggest that people of different ages may differ in their trust assessment strategies, it is likely that the specific effect of age on trust varies according to the context (Steinke, Fritsch, & Silbermann, 2012).

### 6.1.3 Personality traits

A small amount of research has examined general personality traits (openness, conscientiousness, extraversion, agreeableness, neuroticism, intuitiveness) which may correlate with an individual’s overall tendency to trust. Research has shown that individuals with a high propensity to trust are more likely to trust reliable systems than those with lower trust propensity, but their trust may decay more markedly following automation errors (Merritt & Ilgen, 2008). Parasuraman and Riley (1997) found that an operator’s overall propensity to trust automation was distinct from the operator’s trust in a specific system.

Positive correlations between trust levels and intuitive, extraverted and emotional stability personality traits have been demonstrated (McBride, Carter, & Ntuen, 2012; Merritt & Ilgen, 2008; Szalma & Taylor, 2011). However, variability in the consistency of relationships between personality traits and automation across studies suggests that effects of personality on trust may vary as a function of the specific automation and task (Szalma & Taylor, 2011).

### 6.1.4 States

The influence of dynamic human states such as fatigue, stress, mood, and attentional control have been relatively well examined in relation to human-automation interaction, in particular, their effect on workload and performance. However less evidence exists specifically focusing on trust in automation. Attentional capacity often depends on operator workload, but can also be influenced by motivation, tiredness, stress, or boredom (e.g., Reichenbach, Onnasch, & Manzey, 2011).

Several studies have shown that operators with lower attentional control rely more greatly on automated systems than those with higher attentional control, even when automation reliability is low (e.g., Chen & Terrence, 2009). Studies also suggest mood may influence trust development, with happiness significantly increasing trust in an automated system (Merritt, 2011; Stokes, Lyons, Littlejohn, Natarian, Case, & Speranza, 2010). However, while positive attitudes or emotions influence trust in a system, they may result in overreliance (Bailey, Scerbo, Freeman, Mikulka, & Scott, 2006; Merritt, Heimbaugh, LaChapell, & Lee, 2013).

### 6.1.5 Self-confidence

Both trust and self-confidence have been shown to be significant determinants of automation use (Riley, 1996). Early studies of self-confidence in relation to trust in automation suggested a simple relationship: when trust exceeded self-confidence then automation would be utilised, however when self-confidence exceeded trust, manual control would be used (Lee & Moray, 1994; de Vries,
Midden, & Bouwhuis, 2003). However, this has been demonstrated not to be consistent at higher degrees of automation; instead trust has been shown to be influenced by system properties and self-confidence by operator experiences (Moray, Inagaki, & Itoh, 2000).

### 6.1.6 Understanding

Balfe, Sharples, and Wilson (2018), in an observational study of rail signalling operators, found that understanding of the automation was the strongest factor influencing trust, with understanding having a greater effect than the reliability and competence of the automation. They propose that this may indicate important differences between trust examined in real-world systems and that studied in laboratory settings.

### 6.1.7 Subject matter expertise

Subject matter experience refers to expertise in a specific knowledge domain (not in a specific type of automaton). Subject matter expertise often results from extensive experience, leading to greater self-confidence. Individuals with greater subject matter expertise have been shown to be less likely to rely on automation than those with less expertise (Fan, Oh, McNeese, Yen, Cuevas, Strater, Endsley, 2008; Sanchez, Rogers, Fisk, & Rovira, 2014).

### 6.2 Automation factors

The extent to which an automated system successfully performs the task it is intended to achieve has a large effect on automation trust (Adams et al., 2003). The most significant factors of automation use are reliability and the effect of system faults and errors (Lewis, Sycara, & Walker, 2018).

#### 6.2.1 Reliability

Reliability refers to error-rate within an automation system (e.g., the rate at which targets are misidentified). A large amount of evidence demonstrates that automation reliability is related to trust, and that automation which consistently performs well is more likely to be trusted than that which does not (e.g., Kelly et al., 2001; Moray & Inagaki, 1999; Riley, 1989; Parasuraman & Manzey, 2010; Parasuraman & Riley, 1997; Parasuraman, Sheridan, & Wickens, 2008). Sheridan (1998) proposes that operators generalise their future performance expectancies from past experiences, becoming ‘conditioned’ to trust. Bailey and Scerbo (2007) demonstrated that increasing automation reliability results in both increased trust and decreased monitoring. Complementing these results, Moray, Inagaki, and Itoh (2000) manipulated automation reliability in a simulated control system, and they found that operator trust was highly correlated with the reliability of the system, with trust declining over time in line with decreasing automation reliability. Some evidence suggests that it is only the most recent experiences which affect trust in automation (Lee & Moray, 1992).
Operator perceptions regarding the reason for automation unreliability have also demonstrated effects on trust. Bisantz and Seong (2001) conducted an experiment using a decision aid in which one group of participants were informed that automation failures were common, another group were told that the automation was a focus for enemy sabotage, and a third group were given no reliability information. Despite the automation having the same reliability for all groups, the sabotage group displayed less trust in the automation than the control group.

6.2.2 Faults

System faults refer to specific system events, rather than overall automation reliability, and are typically more drastic than errors. A number of studies have demonstrated that, in general, system faults have a negative impact on trust in automation (Adams et al., 2003). When faults occur, trust levels often fall dramatically, and recover much more slowly, even when the automation generally performs reliably (e.g., Kelly et al., 2001; Lee & Moray, 1992; Moray, Hiskes, Lee, & Muir, 1995). System faults also have a greater effect on trust level changes than system successes (Yu, Berkovsky, Taib, Conway, Zhou, & Chen, 2017). The impact of automation faults seems to impact trust and performance separately. After a system fault, trust is much slower to increase than performance, and often does not return to previous levels (Lee & Moray, 1992). In general, large faults have a more negative effect on trust levels than small faults, irrespective of automation performance (Lee & Moray, 1992). However, small faults which vary in magnitude have been shown to diminish trust more than constant large errors, possibly because operators are able to compensate more effectively for systematic errors than for variable errors (Muir & Moray, 1996).

Faults in one element of a control automation have been shown to affect levels of trust in other elements controlled by the same automation, irrespective of the actual performance of the other elements (Muir & Moray, 1996). However, this lowering of trust was shown not to be applied to other similar but independent automation systems.

6.2.3 Predictability

Greater operator familiarity with an automated system can decrease the rate of trust diminishment as a result of errors or faults (Yu et al., 2017), and when given prior knowledge about faults, subsequent faults do not necessarily diminish trust (Riley, 1996). This may be because prior knowledge about potential automation failures reduces the level of uncertainty and risk (Lewis, Sycara, & Walker, 2018).

6.2.4 Transparency and interface

Transparency refers to the extent to which the actions of the automation are understandable and predictable (Endsley, 2017). Automated systems which clarify their reasoning are more likely to be trusted (Simpson, Brander, & Portsdown, 1995; Hoff & Bashir, 2015). Development of trust in automation has been shown to be influenced by the availability of information (Bitan & Meyer, 2007), the amount of feedback (Muir & Moray, 1996), and the accuracy of feedback (Sharples et al., 2007).
6.2.5 Stages and levels of autonomy

Degrees of autonomy (stages and levels) may complicate the development and change in trust level in automation. For example, results of a study investigating the relationship between trust, self-confidence, and automation use at a relatively simple level of automation were not replicated when the same study was conducted using a higher level of automation (Moray, Inagaki, & Itoh, 2000). Higher levels of automation are more complex and therefore less understandable to the operator, which may result in lower levels of trust (Lewis, Sycara, & Walker, 2018). This is supported by a study in which automated target recognition devices were operated at different levels of automation and reliability (Ho, Pavlovic, Myers, & Arrabito, 2013). Participants reported greater levels of trust in the lower automation condition, regardless of automation reliability. A level of automation perceived by the operator as ‘good’ results in greater trust than one perceived as ‘ambiguous’ or ‘poor’ (Merritt et al., 2013).

In an investigation on the extent to which performance of automated decision aids were dependent on the degree of automation, Manzey, Reichenbach, and Onnasch (2012) found that over-reliance on the decision aid emerged independent of the degree of automation, although neglect of automation verification was primarily associated with the highest level of automated aid.

6.2.6 Attributes

Specific physical or other automation attributes have been shown to influence trust levels. For example, the type, size, behaviour, and proximity of a robot has been shown to influence trust in it (Bainbridge, Hart, Kim, & Scassellati, 2008; Tsui, Desai, & Yanco, 2010). The inclusion of some anthropomorphic features such as faces has been shown to increase trust in autonomous systems (Pak, Fink, Price, Bass, & Sturre, 2012). Studies have also demonstrated that people place higher levels of trust in automation which is portrayed as an ‘expert’ system (de Vries & Midden, 2008), although this trust may diminish more quickly when systems make errors (Madhavan & Weigmann, 2005). Looije, Neerinex, and Cnossen (2010) suggest that individuals show a preference for automation which can learn and respond to personality differences.

6.3 Environmental factors

The effect of environmental characteristics on trust in automation have been examined in relation to general automation use, but less so specifically in relation to trust.

6.3.1 Risk

Risk is perhaps one of the most significant environmental factors impacting on trust in automation (Adams et al., 2003). In aviation, for example, trust has been suggested to be an important element in physical terrain alerting systems for pilots, due to the high-risk element associated with flight safety consequences (Borst, Mulder, & van Paassen, 2010). Reliance on automation is modulated by the level of risk inherent in an interaction (Lewis, Sycara, & Walker, 2018). If negative consequences are more likely, operators are more reluctant to use automation and, once trust has
been diminished, operators take longer to re-utilise the automation in high-risk compared to low-risk situations (Riley, 1996). However, advance information regarding system behaviour may alter the perception of risk. When operators know when and how the automation might fail, their trust is not diminished, and reliance remains high (Riley, 1996). Users of GPS route-planners have demonstrated less trust in advice provided when it produced routes which included the presence of driving hazards (Perkins, Miller, Hashemi, & Burns, 2010). However, in comparison to advice from other humans, participants in a different decision-making task demonstrated a preference to rely on automated assistance when making high-risk decisions (Lyons & Stokes, 2012).

6.3.2 Workload and task

An operator’s workload can determine the amount of time and attention is available to monitor an automated system. Task workload has been shown to affect both self-reported trust (Biros, Daly, & Gunsch, 2004; Wetzel, Sheffert, & Backs, 2004; Willems & Heiney, 2002) and reliance behaviours (Daly, 2002; McBride, Rogers, & Fisk, 2011; Rajoanah, Tricot, Anceaux, & Millot, 2008). Positive correlations between trust and reliance have been shown to be moderated by workload, suggesting that operators rely on automation more to complete tasks when workloads are high, regardless of their level of trust (Biros et al., 2004). This was supported by the results of a meta-analysis by Wickens and Dixon (2007), who found that the presence of competing tasks resulted in increased trust in automation.

Environmental distractors have shown mixed effects on trust, with studies showing positive, negative, and no effect on trust from auditory, visual, and spatial distractors (Phillips & Madhavan, 2011; Lees & Lee, 2007). This may be due to the degree to which the distractor interferes with ability to monitor the system, with distractors which mask errors leading to perception of greater automation reliability, and therefore increasing trust (Hoff & Bashir, 2015).

6.3.3 Organisation

Organisational norms and expectations can influence trust in automation. For example, Workman (2005) demonstrated poor attitudes combined with high subjective norms increased a tendency to misuse (over-trust) decision support automation.

6.4 Learned trust – Dynamic trust changes over time

Previous knowledge, past experiences, and current interactions dynamically influence trust in automation over time (Hoff & Bashir, 2015). Desai et al. (2013) found that trust in robots was affected more by early failure than by later reductions in reliability. Dzindolet, Peterson, Pomranky, Pierce, and Beck (2003) examined trust in an automated decision aid. Operators initially displayed a positive bias towards the automation, expecting it to perform in a reliable and trustworthy manner. After an experience of automation failure, trust decreased sharply, and operators were reluctant to trust even reliable aids. On receiving an explanation for the occurrence of errors, participants displayed increased trust and reliance in the decision aid, even when performance levels suggested the trust was not warranted.
Trust in highly reliable automation which has been established over extended time periods can result in out-of-the-loop performance problems and skill decay if manual control is reallocated (Hillburn, Molloy, Wong, & Parasuraman, 1993; Parasuraman, Mouloua, & Molloy, 1996).
7 Measuring trust in automation

7.1 Issues measuring trust in automation

As a hypothetical latent construct which cannot be directly observed or measured but only inferred, measuring trust experimentally is a difficult task. In this way, trust is similar to other constructs such as workload and mental models (Muir, 1994). Billings, Schaefer, Chen, and Hancock (2012) note a number of issues still to be addressed in order to satisfactorily quantify dynamic trust in automation: defining what is being measured, defining when it should be measured, and defining how it should be measured.

What is being measured is still highly varied within current literature. As discussed previously, differences exist in conceptualising trust as a belief, attitude, willingness, or behaviour. Different types of trust are also examined: trust propensity, dispositional trust, history-based trust, experience-based trust, initial trust, ongoing trust, affective trust, cognitive trust, and so on. Differences in results may also occur as a result of the way participants are asked to rate their subjective level of trust. For example, “to what extent do you trust the automation”, “to what extent do you believe the automation is trustworthy”, “to what extent do you trust the automation to successfully complete the task” and “to what extent has your trust in the automation changed?” probably do not measure the same trust construct. The term ‘trust’ has also been found to be problematic for particular study participants. For example, Adams et al. (2003) noted that, among the military personnel with whom they were working, ‘trust’ was viewed as too “touchy-feely” to be suitable for describing working relationships with either automation or other crew members, with ‘dependability’ and ‘reliability’ being offered as more suitable terms.

When trust is measured within a study has changed over time. Early studies often only measured trust once, at the end of a session. Pre- and post- levels of trust were sometimes assessed. In current studies focusing exclusively on trust (rather than as one of a number of factors contributing to automation use), it is more common for trust to be measured multiple times throughout a session, often at the end of each trial/set of trials, or at specified time intervals. When measured over time, trust is often assessed as a loss or gain in trust, rather than at a specific trust level (Alarcon, 2017).

How trust is measured possibly represents the largest dichotomy within trust measurement. Of the studies examined for this literature review, nearly two thirds used some form of subjective self-assessment rating of trust using a scale, while approximately one quarter of studies used a behavioural outcome to infer trust. Comparing results across these studies can be problematic. Justification for choice of measurement format is often not given. Subjective scales have the advantage of ease of use, and provide a possible means to access the internal cognitive and affective processes contributing to trust which may not be evident from resultant behaviour. However, just over half of the studies reviewed which measure trust in this way created their own trust scale, leading to the difficulties with discrepancies in trust definitions described above. Desai,
Kaniarrasu, Medvedev, Steinfeld, and Yanco (2013) also demonstrated that trust questionnaires disseminated at the end of a study do not reflect real-time trust in robots. Additionally, and as has been well-discussed in relation to self-report scales, participants may not be able or willing to accurately report their true attitude. User trust in automation has been demonstrated to consist of a combination of explicit and implicit attitudes, with users being incapable of describing the effects of implicit attitudes on trust (Merritt, Heimbaugh, LaChapell, & Lee, 2013).

Behavioural trust measures have theoretical basis in the models which describe the relationship between attitudinal trust and behavioural reliance (or cooperation) on automation (e.g., Lee & See, 2004). They are intended to indirectly assess levels of trust. Behavioural measures have the advantage of providing a more potentially consistent method of measuring trust, and they can more easily be used as a basis for modelling and prediction (Drnec, Marathe, Lukos, & Metcalfe, 2016). However, it is often difficult to separate the effects of other factors such as workload, stress, tiredness, or workplace directives on behaviours in order to examine the specific effect of trust (Drnec et al., 2016). Dekker and Hollnagel (2004) have also argued for a need for greater precision in defining behavioural constructs such as ‘complacency’ in order to avoid construct creep and ‘folk models’.

Bindewald, Rusnock, and Miller (2018) produced a well-considered paradigm for examining how trust-based behaviours are manifested in human-machine teams which argues that trust is not the objective for designers of automation. Rather, trust should be viewed solely as a means to an end – that of influencing human behaviour – and therefore behavioural measurements have more validity than self-report. While probably true within their focus on air force automation use, this view of trust ignores the human element inherent within automation interaction, and it fails to capture the “willingness” aspect of trust. A human operator can be required to comply with automation in which he or she does not trust. Drnec at al. (2016) argue that re-framing the entire field of trust in automation to one of reliance and compliance would remove the need for making assumptions and drawing inferences about the manifestations of subjective constructs such as trust. However, focusing purely on behavioural reliance and compliance in automation has the potential of reducing humans to little more than cogs within the automation machine. It is interesting to note that Lee (2008) has begun to focus on outcomes of automation for performance, safety and satisfaction. As the field matures, it is possible to envisage a greater use of behavioural reliance and compliance measures to assess performance and safety, with the addition of measures of ‘confidence’ or ‘trust’ to capture human willingness and satisfaction in use of automation.

A very small number of studies are beginning to use multiple measures of trust (e.g., de Vries, Midden, Bouwuis, 2003; de Visser, Monfort, McKendrick, Smith, McKnight, Krueger, & Parasuraman, 2016; Gold, Körber, Hohenberger, Lechner, & Bengler, 2015). This may help support future refinement of both behavioural and self-report measures to better capture trust. However, Hergerth, Lorenz, Krems, and Toenert (2015) found that self-report and behavioural measures of trust in highly automated driving were not correlated, indicating that they may in fact be measuring different things.

Some studies are beginning to examine the utility of measuring potential physiological and ‘neuroergonomic’ correlates (Borghetti, Giametta, & Rusnock, 2017). These are still at early stages,
but may prove beneficial in effectively capturing trust assessments, particularly dynamic changes in trust. However, in order to validate them, it is still necessary to calibrate such measures against some measure of subjective trust (Bindewald, Rusnock, & Miller, 2018). It is still unclear at this stage the extent to which, in using these types of measures, effects due to trust can be distinguished from effects due to other factors, particularly workload.

7.2 Current trust in automation measures

7.2.1 Self-report scales

Self-report scales produced specifically for a study tend to consist of between one and 10 items. Scales usually range from indicators of “not at all” to “completely [trust]”. An odd number of scale points are more common, which allows participants to record a neutral trust level.

By far the most commonly used validated scale is that developed by Jian, Bisantz, and Drury (2000). This 7-point scale consists 12 items which are intended to measure global trust in automation. Items include: the system is dependable, and I can trust the system. A recent study utilising the scale showed excellent internal reliability (Cronbach’s alpha = .93) (Buckley, Kaye, & Pradhan, 2018).

Other scales include Mayer and Davis’s (1999) propensity to trust scale, Lee and Moray’s (1994) subjective rating scale, the Human Computer Trust questionnaire (Madsen & Gregor, 2000), and a cross-cultural trust in automation scale (Chien et al., 2014). Scales are also being developed to measure trust in particular types of automation, such as autonomous vehicles (Garcia, Kreutzer, Badillo-Urquiola, & Mouloua, 2015), and robotics (Yagoda & Gillan, 2012).

Most scales are in written form, although some are completed verbally. Desai, Kaniarrasu, Medvedev, Steinfeld, and Yanco (2013) used buttons on a robot’s controller to allow a user to indicate trust increases and decreases at specified times when prompted by a light on the control (see Figure 21).

Figure 21. Robot controller with buttons to indicate trust level changes (from Desai et al., 2013).
7.2.2 Behavioural correlates

Common behavioural measures of trust in automation are outlined in Table 4.

Table 4. Common behavioural measures of trust in automation

<table>
<thead>
<tr>
<th>Behaviour</th>
<th>Example studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choice of manual or automatic task completion (automation disuse)</td>
<td>de Vries, Midden, and Bouwhuis (2003)</td>
</tr>
<tr>
<td>Level of automation chosen by the operator in adaptive automation</td>
<td>Desai et al. (2013)</td>
</tr>
<tr>
<td>Reaction time: higher trust equating to longer reaction times</td>
<td>Payre, Cestac, and Delhomme (2016)</td>
</tr>
<tr>
<td>Reliance behaviour:</td>
<td></td>
</tr>
<tr>
<td>Acceptance of non-indication of advice given or action needed</td>
<td>Bindewald, Rusnock, &amp; Miller (2018)</td>
</tr>
<tr>
<td>Hand-over and take-back of automation control</td>
<td>Gao, Lee, and Zhang (2006)</td>
</tr>
<tr>
<td>Monitoring</td>
<td>Gremillion et al. (2016)</td>
</tr>
<tr>
<td>Compliance behaviour:</td>
<td></td>
</tr>
<tr>
<td>Acceptance of advice given, or action chosen, by automation</td>
<td>Bindewald, Rusnock, &amp; Miller (2018)</td>
</tr>
<tr>
<td>Checking</td>
<td>de Visser et al. (2017)</td>
</tr>
<tr>
<td>Confirmation: withdrawing own decision to comply with automation decision</td>
<td>Antifakos et al. (2005)</td>
</tr>
<tr>
<td></td>
<td>Gaudiello, Zibetti, Lefort, Chetouani, and Ivaldi (2016)</td>
</tr>
<tr>
<td>Startle response</td>
<td>Waytz, Heafner, and Epley (2014)</td>
</tr>
</tbody>
</table>

7.2.3 Physiological and neural measures

Physiological, neural and combined measures of trust used in automation trust studies are outlined in Table 5.

Table 5. Other measures of trust in automation

<table>
<thead>
<tr>
<th>Measure</th>
<th>Example studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaze behaviour through eye tracking</td>
<td>Gold, Körber, Hohenberger, Lechner, and Bengler (2015)</td>
</tr>
<tr>
<td></td>
<td>Hergerth, Lorenz, Vilimek, and Krems (2016)</td>
</tr>
<tr>
<td></td>
<td>Johnson, Duda, Sheridan, &amp; Oman (2017; to examine attention and situation awareness rather than trust)</td>
</tr>
<tr>
<td>Heart-rate beat and variability: lower heart-rate presumed to indicate higher trust level</td>
<td>Peterson, Robert, and Yang (2017)</td>
</tr>
<tr>
<td></td>
<td>Khalid, Shiung, Nooralishahi, Rasool, Helander, Kiong, and Ai-vryn (2016)</td>
</tr>
<tr>
<td>EEG: to examine neural correlates</td>
<td>Devlin (2013)</td>
</tr>
<tr>
<td></td>
<td>Nuamah, Oh, and Seong (2015; proposal only; to measure intuitive vs analytic trust)</td>
</tr>
<tr>
<td></td>
<td>Borghetti, Giametta, and Rusnock (2017; to examine continuous operator workload rather than trust)</td>
</tr>
<tr>
<td>Method</td>
<td>Reference</td>
</tr>
<tr>
<td>---------------------------------------------</td>
<td>---------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Oxytocin administration to influence trust and compliance</td>
<td>Anumandla, Bray, Thibeault, Hoang, Dascalu, Harris, and Goodman (2011) de Visser et al. (2017)</td>
</tr>
<tr>
<td>Cerebral blood flow velocity and oculomotor fatigue</td>
<td>Funke, Warm, Matthews, Funke, Chiu, Shaw, &amp; Greenlee (2017)</td>
</tr>
<tr>
<td>fMRI: to examine brain activation patterns</td>
<td>Drnex, Marathe, Lukos and Metcalfe (2016) Goodyear, Parasuraman, Chernyak, de Visser, Madhavan, Deshpande, &amp; Krueger (2017)</td>
</tr>
<tr>
<td>fNIRS: to examine areas of brain activation</td>
<td>Heironimus, Guznov, and Pfahler (2017)</td>
</tr>
<tr>
<td>Affective computing analysis</td>
<td>D’Mello, Kappas, and Gratch (2017)</td>
</tr>
<tr>
<td>Facial and voice tracking</td>
<td>Khalid, Shiung, Nooralishahi, Rasool, Helander, Kiong, and Ai-vryn (2016)</td>
</tr>
</tbody>
</table>

More generally, Lottridge, Chignell, and Jovicic (2011) consider a number of different methods for measuring emotion, which may have some utility when considering potential trust measurements. Additionally, a yet-to-be-released study by Oglesby (submitted) which examines measures and metrics of human-automation systems safety and performance may provide further insights into effective trust measurements.
The body of research examining trust in automation is expanding rapidly, particularly in the fields of semi-autonomous cars, decision aids, unmanned autonomous vehicles, and robotics. While commonly-used definitions, models, and measures of trust have emerged, lack of clarity and consistency remains, particularly in regards to defining and measuring trust. This creates difficulty in synthesising results across the literature to confidently build upon previous work. Ongoing debate in this area considers the merit of measuring reliance and compliance behaviours as an alternative to subjective trust measures against the negative consequences of removing measures of human emotion and attitude from consideration. Debate also surrounds the most commonly used automation taxonomy, which categorises automation according to levels and stages of automation. Alternatives propose placing greater emphasis on human-automation collaboration. However, a number of prominent researchers in the trust-in-automation field argue that discarding a well-established taxonomy has the potential to create confusion across autonomy research, design, and construction. As what is technically possible continues to advance, human interactions with autonomous systems will increasingly involve teamwork relationships, requiring the consideration of human-automation cooperative behaviour. The role played by trust in such relationships remains of consequence in ensuring that humans continue to successfully utilise autonomous systems in order to extend human performance in safe and fulfilling ways.


https://doi.org/10.1109/HRI.2013.6483596


https://doi.org/10.1002/hfm.4530050105


Kelly, C., Boardman, M., Goillau, P. J., & Jeannot E. (2003). Guidelines for trust in future ATM systems: A literature review. (HRS/HSP-005-GUI-01) Eurocontrol. Retrieved from https://www.researchgate.net/publication/311065869_Guidelines_for_Trust_in_Future_ATM_Systems_A_Literature_Review?enrichId=greq-2236f387dd0d069cccf31031a23399774-XXX&enrichSource=Y292ZXJQYWhdOzMxMTA2NTg2OTtBUzo0MzM1NDE5NTUtQ1NDRAMTQ4MDM3NTY3NDU0Mw%3D%3D&el=1_x_2&_esc=publicationCoverPdf


Society of Automotive Engineers (SAE) International (2014). *Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles* (J3016_201609). Retrieved from https://www.sae.org/standards/content/j3016_201609/


https://www.researchgate.net/publication/235171265_Decision_Support_Automation_Research_in_the_En_Route_Air_Traffic_Control_Environment


