Calibrating Trust Between Humans and Artificial Intelligence Systems

Submitted for the degree of Doctor of Philosophy (PhD)

Martin Ingram, MSc, MA

School of Psychology,
College of Science and Engineering,
University of Glasgow

June 2022

Abstract

As machines become increasingly more intelligent, they become more capable of operating with greater degrees of independence from their users. However, appropriate use of these autonomous systems is dependent on appropriate trust from their users. A lack of trust towards an autonomous system will likely lead to the user doubting the capabilities of the system, potentially to the point of disuse. Conversely, too much trust in a system may lead to the user overestimating the capabilities of the system, and potentially result in errors which could have been avoided with appropriate supervision. Thus, appropriate trust is trust which is calibrated to reflect the true performance capabilities of the system. The calibration of trust towards autonomous systems is an area of research of increasing popularity, as more and more intelligent machines are introduced to modern workplaces.

This thesis contains three studies which examine trust towards autonomous technologies. In our first study, in Chapter 2, we used qualitative research methods to explore how participants characterise their trust towards different online technologies. In focus groups, participants discussed a variety of factors which they believed were important when using digital services. We had a particular interest in how they perceived social media platforms, as these services rely upon users continued sharing of their personal information. In our second study, in Chapter 3, using our initial findings we created a human-computer interaction experiment, where participants collaborated with an Autonomous Image Classifier System. In this experiment, we were able to examine the ways that participants placed trust in the classifier during different types of system performance. We also investigated whether users' trust could be better calibrated by providing different displays of System Confidence Information, to help convey the system's decision making. In our final study, in Chapter 4, we built directly upon the findings of Chapter 3, by creating an updated version of our human-computer interaction experiment. We provided participants with another cue of system decision making, Gradient-weighted Class Activation Mapping, and investigated whether this cue could promote greater trust towards the classifier. Additionally, we examined whether these cues can improve participants' subjective understanding of the system's decision making, as a way of exploring how to improve the interpretability of these systems.

This research contributes to our current understanding of calibrating users' trust towards autonomous systems, and may be particularly useful when designing Autonomous Image Classifier Systems. While our results were inconclusive, we did find some support for users preferring the more complicated interfaces we provided. Users also reported greater understanding of the classifier's decision making when provided with the Gradient-

weighted Class Activation Mapping cue. Further research may clarify whether this cue is an appropriate method of visualising the decision-making of Autonomous Image Classifier Systems in real-world settings.

Table of Contents

List of Tables

List of Figures

Definitions

AI – Artificial Intelligence

ART-ANOVA – Aligned Rank Transform Analysis of Variance

AICS – Autonomous Image Classifier System

CAM – Gradient-weighted Class Activation Mapping

CDSS – Clinical Decision Support Systems

FoMO – Fear of Missing Out

GUI – Graphical User Interface

LOA – Levels of Automation

NASA-TLX – NASA Taskload Inventory

OIDV4 – Open Images Database V4

OIDV5 – Open Images Database V5

PAS – Perfect Automation Schema

SAT – Situation awareness-based Agent Transparency model

SCI – System Confidence Information

SCI+CAM – System Confidence Information and Gradient-weighted Class Activation Mapping (Interface)

SMS – Social Media Services

PTMQ – Propensity to Trust Machines Questionnaire

TAM – Technology Acceptance Model

UAV – Unmanned Aerial Vehicles

Acknowledgements

I would like to take this opportunity to thank the people who have helped me see this project through to its completion.

My supervisor Frank, who has always been available with helpful advice, support and encouragement.

My lab group, especially Cristina and Greta, who were always around to share advice and/or complaints (mostly complaints) over coffee or tea.

The folks at my industrial sponsors, Qumodo, especially, Ben, Sophie, and Diana who have always been supportive and enthusiastic about my research.

My friends away from university, including but not limited to: Ross, Vanessa, Jen, Mark, Sam and Chris, with whom I've not spent nearly enough time with over these years.

My parents, who I have also not seen nearly enough of over the last couple of years – *perhaps there is a pattern there…*

And lastly, to my partner Amalia, who has been with me each step of the way, and if anything has seen too much of me during this journey. Your patience and encouragement have kept me going through the good times and the bad.

This work was supported by the Economic and Social Research Council and the Scottish Graduate School of Social Science: Reference No. ES/P000681/1.

Author's declaration

University of Glasgow

College of Science and Engineering

Statement of Originality to Accompany Thesis Submission
Name: Martin Ingram

Registration Number: XXXXXXXX

I certify that the thesis presented here for examination for a PhD degree of the University of Glasgow is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it) and that the thesis has not been edited by a third party beyond what is permitted by the University's PGR Code of Practice.

I declare that the thesis does not include work forming part of a thesis presented successfully for another degree.

I declare that this thesis has been produced in accordance with the University of Glasgow's Code of Good Practice in Research.

I acknowledge that if any issues are raised regarding good research practice based on review of the thesis, the examination may be postponed pending the outcome of any investigation of the issues.

Signature: _____ Date: **20/06/2022**

This completed statement must be bound into the submitted copies of the soft-bound thesis.

Chapter 1: Trust and Autonomous Technologies

1.1.1 Aims and Overview

This literature review has aimed to provide an overview of the existing literature examining trust towards autonomous systems. The research within this thesis was intended to inform the design of an autonomous image classifier system. In carrying out this literature review, research involving similar types of autonomous systems, such as diagnostic aids, was considered as more relevant than research involving autonomous vehicles or social robotics. Nonetheless, some research from these both of these fields was still included in our review, particularly where there were findings that appeared generalisable to exploring trust towards an autonomous image classifier. We attempted to ground the experiments within this thesis within existing models of trust towards automation, as a way of exploring the many different factors that could influence trust. The models that informed our research are some of the most influential within the field, namely: Lee and See (2004) and Hoff and Bashir (2015), and de Visser and colleagues (2018). These models were used as a starting point in developing our initial understanding the literature, and were useful in identifying other relevant research by looking at the studies cited within these models, and also by searching for newer research that cited these models. Each chapter within this thesis introduces a new concept with which we explore trust towards technology. For example, Chapter 2 examines social media platforms, Chapter 3 examines System Confidence Information, and Chapter 4 examines Gradient-weighted Class Activation Mapping. Additional research for each concept was sought and reviewed for each chapter as they were written.

Advances in computing science, such as Deep Learning and Convolutional Neural Networks have enabled new technologies to undertake more complicated tasks. Autonomous systems can now be trained to make complex decisions and can often work uninterrupted to process large workloads, all with reduced supervision from human users. The flexibility of these systems also permits a wide range of potential applications, where intelligent machines can augment human workers to improve productivity within many modern workplaces. For example, McKinney and colleagues (2020) report on the use of an autonomous system that was trained to identify breast cancer from mammograms. The system had an accuracy comparable to human oncology experts and if used correctly, they predicted it could lower the specialists' workload by up to 88% (McKinney et al., 2020). As intelligent technologies become more pervasive there has also been a growing interest in how much trust users should place in these systems. Ideally, users will not place too

little trust in a well-functioning machine, nor will they place too much trust in a faulty machine (Muir, 1987; Parasuraman & Riley, 1997). Instead, users' trust should be calibrated to reflect the system's true performance capabilities, thereby ensuring appropriate trust towards the autonomous system. This calibration of trust towards automation is a rapidly growing area of research and has recently been advanced by the development of newer, more accessible autonomous technologies.

1.1.2 The History of the Trust-in-Automation Field

In many ways, research into human-computer teams has been limited by the computational power of the technology available. While there has been interest in man-machine interactions since the 1960's (Neurath et al., 1969), the research, and by extension our understanding of man-machine interactions, has been limited by a lack of intelligent technology that can be applied to a variety of workplaces and environments. To this extent, early research tended to use technology which flowed from areas where autonomous technologies are most readily available, such as in military domains and the aviation industry (Biros et al., 2004; Bragg et al., 1998; Chen & Terrence, 2009; McGuirl & Sarter, 2006; McGuirl et al., 2009). While it does not appear to be the case for all research carried out in the early days of this field, Pak and colleagues (2017) argue that it is important to consider that many of these studies would have involved participants who are highly trained and/or educated. Moreover, these studies also typically involve individuals who are part of a highly specific culture, particularly those from military backgrounds, where the organisational culture may have a direct influence on their attitudes towards trust and teamwork (Pak et al., 2017). While this does not necessarily invalidate this research, it is something that is worth considering in comparison to research involving newer technologies, which may involve more generalised participant populations.

In more recent years, while a significant number of studies still involve military personnel and/or technology (Lyons et al., 2016 Ho et al., 2017; Rogers et al., 2019; Selkowitz et al., 2017; Wright et al., 2020), a greater amount of research has been carried out using more diverse technologies and participant pools. These include investigations involving trust towards robots (Desai et al., 2013; Maurtua et al., 2017; Hancock et al., 2011); fault detection systems (Chavaillaz and Sauer, 2017; Yu et al., 2019), automated baggage scanners (Merritt et al., 2013; Chavaillaz et al., 2019) and decision support systems (Goddard et al., 2014; Lyell et al., 2018; de Visser et al., 2014). However, out of all the research currently investigating trust towards automation, the most significant contributor appears to be the automobile industry, in their pursuit of fully autonomous

vehicles (Pammer et al., 2021; Verberne et al., 2012; Verberne et al., 2015; Niu et al., 2018; Jing et al., 2020; Koo et al., 2015).

This interest in autonomous vehicles is understandable, given the ubiquity of the car as a mode of transport, and the considerable amount of road accident fatalities each year. As a result, car manufacturers are targeting autonomous vehicles as a realistic way to improve road safety (Fagnant and Kockelman 2015). Yet while car manufacturers may be beginning to push the development of autonomous technology, the success of these vehicles is still largely dependent on whether human drivers will accept them, and support for this acceptance is currently inconclusive. Abraham and colleagues (2016) surveyed drivers across a wide range of age groups and asked how willing they were to use vehicles with varying capacities of automation. Responses indicated that drivers were more favourable towards vehicles with partial and full autonomy capability, as opposed to those with no autonomous capability, suggesting support from drivers for autonomous vehicles. However, when asked about the features and types of autonomy they would be willing to use, drivers showed a higher overall preference for automation that could reduce and mitigate collisions, yet they were less favourable towards automation that assisted with control of driving (Abraham et al., 2016). This could mean that while there is an appetite for the introduction of autonomous vehicles, drivers are not prepared to relinquish full driving control. This is perhaps understandable given the novelty of this technology. This may also help to explain why there is so much research currently being done involving trust towards autonomous vehicles. Nonetheless, autonomous vehicles are no exception, and it is likely that most autonomous systems will require appropriate trust from their users. However, ‘Autonomous Systems’ are something of an umbrella term, and many different types of machines can be captured underneath this phrase.

1.2 Defining Autonomous Systems

Broadly speaking, autonomous systems are machines that have been designed to complete tasks which would have been previously carried out by humans (Parasuraman and Riley, 1997). By doing so, these systems can liberate their human users from repetitive, labour-intensive tasks, and instead allow them to undertake more complex, supervisory duties instead. Arguably the clearest example of this can be seen in the agriculture industry, where much of the labour in modern farming is now carried out by machines. Historically, human workers would have spent their entire working day harvesting crops within fields, often within harsh weather conditions, with relatively limited productivity. Beginning with the original ‘Reaper’ machine designed by Scotsman Patrick Bell in 1826, and culminating in the modern day Combine Harvester, automation

has greatly improved the productivity of farming (West, 1967). With more widespread availability of newer, smarter technologies, many other modern workplaces are also expected to undergo rapid, technology-induced change, known as the Industrial Revolution 4.0 (Morrar et al., 2017). Through advances in Artificial Intelligence (AI), such as Deep Learning, autonomous systems can now be trained and retrained to handle newer and more complicated tasks, which gives autonomous systems a flexibility that was not previously available. Yet not all autonomous systems are designed the same, and much of their role within the workplace will be defined by the types of tasks they are designed to undertake.

1.2.1 Levels of Autonomy

When categorising automation, Parasuraman and colleagues (2000) propose that autonomous systems can be separated into two different categories: Information-based Automation and Decision-based Automation. Across these two categories, they also constructed four sub-stages which can reflect the increasing complexity of autonomous systems. For Information-based automation, at the lowest stage (1) the system is expected to be involved in the Acquisition of Information, wherein the system is used to collect data without analysis. More advanced Information-based automation (2) would also be expected to be involved in the Manipulation/Analysis of information, wherein the system collects and then processes the data. Thus, for Information-based automation, the human operators remain responsible for any decision making that is made for completing tasks, and the system is employed as more of a tool by the user. Within decision-based automation (3), the system would be expected to go further, and be able to recommend decisions based on analysis of information. At the highest stage (4), decision-based automation would also be able to execute actions based on these decisions, with human users only playing a limited role in this decision-making.

Within each of these four stages, the automation can be graded on the Levels of Automation (LOA) which illustrate how much of the task the computer handles without human consultation, a concept that was originally proposed by Sheridan & Verplank (1978). For example, a system with Facial Recognition software may have high LOA in stages 1 and 2, if it can collect and analyse visual data from its camera sensors. Yet for the decision-based automation stages (3 & 4), it may have low LOA if it only offers limited recommendations to the human operator, and especially if the user is responsible for making the final decision based on the feedback from the system. Each of these levels directly influence the degree of input a human user has on the human-computer team, with higher levels leading to less human influence. This effectively means as the machine becomes more autonomous the human is left out of more of the decision-making process,

and is thus more reliant on the machine (Parasuraman et al, 2000). When there are potential consequences for human safety, or the ultimate outcome involves lethality, such as in military applications, then trust is likely to be further influenced by the task context (Parasuraman and Wickens 2008; Pak et al., 2017). Therefore, it is also important to consider that trust in human-machine teams may be influenced to a degree by the inherent nature of the task.

1.2.2 The Nature of the Task

With their increasing flexibility, autonomous systems are now being employed in a variety of settings, and with that comes different stakes attached to successes and failures of the system. For example, Autonomous Image Classifier Systems (AICS), which are systems trained to categorise the contents of image data and can be used in both low-stakes and high-stakes settings. Low-stakes applications of AICS include 'Pl@ntNet', an app-based image classifier trained to identify plants and flowers from images (Goëau et al., 2014). In contrast, the AICS used by McKinney and colleagues (2020) is intended to identify breast cancer from mammography images and represents a much higher stakes use of automation. While both examples use similar AICS technology, there are diverging stakes attached to system successes and failures, one may be used by hobbyists to identify plants, while the other may be used to make a critical diagnosis for patient's health. In both instances, continued use of the AICS are dependent on the user trusting the system, yet with the breast cancer classifier example, errors may be met with greater losses in trust from the user, given the consequences of system errors.

Similar research involving autonomous systems used in high stakes situations can be seen with Clinical Decision Support Systems (CDSS), which can help health practitioners to make diagnoses (Goddard et al, 2014; Lyell et al., 2018), automated weapons detection systems (Merritt et al., 2013; Chavaillaz et al., 2019), autonomous squad member robots (Selkowitz et al., 2017; Wright et al., 2020), and Unmanned Aerial Vehicles (UAV) (Lin & Goodrich, 2015; Rogers et al., 2019). Arguably, there is scope for high stake situations involving all autonomous systems at some point in their deployment, yet it may still be difficult to replicate these conditions within controlled laboratory settings. Additionally, while these systems are intended to reduce the workload of human users, if participants can complete the task themselves, it may be difficult to stop them from disusing the autonomous system in laboratory experiments if these high-stake consequences are not present. An interesting way of exploring this was used by Verame and colleagues (2016), who asked participants to work with an automated handwriting detection system in conditions with and without performance-related financial incentives.

They found that participants who were offered financial incentives linked to the performance of the handwriting system were more likely to collaborate with the system. In contrast participants who were offered a fixed reward that was not linked to task performance were more likely undertake the manual tasks, which reportedly required the least effort from participants. This suggests that participants in laboratory-based experiments can be motivated to collaborate more with autonomous technology if offered financial incentives. At the same time, researchers' ability to do so will also be dependent on the budget for the study, and the ethical frameworks that govern the field in which the research is taking place.

Ultimately, research involving trust towards automation has been carried out in a variety of settings, involving many different types of automation completing different tasks, all with participant groups from various backgrounds. Nonetheless, across all these studies researchers have still been able to identify trends and patterns which explain how human users calibrate their trust towards autonomous systems.

1.3 Defining Trust

Trust is an often-amorphous concept that is associated with the intentions and motivations of others, and is essential for cooperative behaviour between individuals, groups, organisations, and other entities (Bhattacharya et al, 1998; Deutsch, 1960; Lewicki et al., 1998; Jones & George, 1998; Rousseau et al., 1998). The nature of trust varies depending on the field examining it: for economists, trust is something to be calculated, whilst sociologists examine trust through the lens of social norms and societal influences, and psychologists focus more on interpersonal trust (Rousseau et al., 1998). At a theoretical level when one person places trust in another person, they do so because they are confident that their interests and wellbeing will be promoted by the other person, without fear of exploitation (Read, 1962; Lewicki et al., 1998). For trust to occur, Jones and George (1998) suggest that both parties need to have confidence in the values and trustworthiness of each other. They will also likely sharin favourable attitudes towards each other, and should both experience positive emotional affect as a result of being within the relationship (Jones & George, 1998).

Across the different fields concerned with examining trust, Rousseau and colleagues (1998) argue the most common factors are (1) the individual's willingness to be vulnerable and (2) their confident expectations in the other person. When trusting others, successful cooperation is only achieved if both parties can orient away from focussing entirely on their own individual interests, and are instead willing to accept some costs to themselves (Deutsch, 1960). When an individual places trust in someone, their decision to

do so will also likely be informed by the trustworthiness of the recipient, i.e. the trustee (Dunn & Schweitzer, 2005). The trustworthiness of an individual is typically based on perceptions of their personal attributes, such as their competence, loyalty, consistency, fairness, and the integrity of their actions and beliefs (Rotter, 1980; Hardin, 2002; Dunn & Schweitzer, 2005, Butler, 1991). Thus, interpersonal trust between humans can be seen as an act of deliberate vulnerability, where one individual depends on another with the optimistic expectation that this will be beneficial to them, but this can only occur if the recipient is trustworthy (Hosmer, 1995; Lewicki et al., 1998; Hardin, 2002; Dunn & Schweitzer, 2005). Much of the ideas from the literature on trust between humans also appears to be relevant when considering trust towards technology.

1.3.1 Trust Towards Technologies

When a human user places trust in an autonomous system, they do so in the belief that the system will successfully perform the tasks that it was designed to do, allowing the user to supervise performance of the task. Across the trust in automation literature, the consensus appears to be that users will be much less likely to trust a technology when they see it make errors and mistakes (Lee and See, 2004; McGuirl & Sarter, 2006; Merritt et al, 2015; Hoff & Bashir, 2015; Yu et al., 2019). While there are many similarities between human-human trust and human-machine trust, trust towards technology is based primarily on the performance of the system. Lee and See (2004) highlight 3 core components within trust towards automation: *Performance*, *Process*, and *Purpose*. *Performance* relates to the reliability, competency and predictability of the system, *Process* relates to how the system achieves this, and whether this is appropriate for the task, while *Purpose* relates to how well the system is being used by the operator, and whether this is within the original parameters defined by the designers (Lee & See, 2004). While aspects of human-human trust can also be also evaluated within these components, there are arguably unique differences in how we evaluate humans versus machines. In comparing human-human and human-machine trust, Madhavan and Wiegmann (2007) contrast the development of trust towards human advisors and automated systems. While many of the components for developing trust between these entities are similar, Madhavan and Wiegmann (2007) argue that machines are expected to perform more consistently across situations, yet can also be invariable and restricted in their approach. By contrast, a competent human advisor would be expected to be more adaptable, and able to change their decisions and behaviours based on situational changes (Madhavan & Wiegmann, 2007). Arguably, newer technologies that make use of Deep Learning are now more adaptable that the decision support systems discussed by Madhavan and Wiegmann (2007). However, even when adapting to new

tasks or changes in their parameters, these systems still require a resource-intensive retraining process, and may still not be as capable as human advisors for relearning and adapting in real-time.

This relationship between trust and performance may be particularly critical when humans work with autonomous systems, compared to when working with other humans or non-automated technology. When Furlough and colleagues (2019) asked participants to ascribe blame when reading about human-robot task failures, the robot was more likely to receive a higher proportion of the blame if it was described as autonomous. Likewise, Berkeley and colleagues (2015) report that when witnessing forecasting errors, human observers are much more likely to lose confidence in forecasting algorithms than human forecasters, even if both make the same mistake. This suggests an innate link between system performance and users' trust towards the system. In contrast, Merritt (2011) suggests that trust towards automation may be formed from both logical and emotional processes, with evaluations based on both how well the system performs, and how much the user likes the machine. Support for this can be seen in work by Thüring and Mahlke (2007), who report that instrumental (functional) and non-instrumental (aesthetic) interface features both shaped participants' positive attitudes towards electronic musical devices. While instrumental features may directly influence users' perceptions of reliability and functionality, non-instrumental aesthetic features also appeared to improve the user's emotional experience of the device. This suggests that trust towards autonomous systems is not entirely based upon logical evaluations of system performance, but rather a combination of different perceptions from the user. Ultimately, for the purpose of this thesis, when measuring trust towards technology this was primarily based on participants' willingness to continue to use, operate, and rely upon the system. This was based on Parasuraman and Riley's (1997) views on distrust and mistrust of autonomous systems, in which optimal trust should still see a user critically evaluate a system even when it is correct, yet not to the extent that they discontinue their use of system. On top of this, other factors relating to the design of the system and interpersonal differences between users were also considered as moderating factors, in line with the models of Lee and See (2004) and Hoff and Bashir (2015).

Indeed, when it comes to the adoption of any new technology, there are a wide range of factors that can influence consumers' judgements, and by extension the ultimate success of the system. Extensive work by Viswanath Venkatesh has sought to map the myriad factors which can inform how users accept new technologies, which are framed in the continuously evolving Technology Acceptance Model (TAM). Earlier versions of TAM, building on the initial model put forward by Davis and colleagues (1989) focussed

on the perceived usefulness and perceived ease of use of the system, as the primary determinants for technology acceptance (Venkatesh & Davis, 2003; Venkatesh et al., 2003). More recently, later iterations of TAM have sought to integrate further, more diverse determinants of acceptance, such as system price, the user's habits, and even the hedonistic pleasure gained from using the system (Venkatesh et al., 2012; Venkatesh, 2015). Technology acceptance is its own distinct field yet trust in automation research can still learn from the many different psychological, social and environmental factors which can inform the user's ultimate acceptance of technology. While the successful adoption of new technologies is tied to users' acceptance of them, the users also need to learn to use the technology correctly. Just because someone accepts a new technology, it does not automatically follow that they will use it appropriately. This is particularly the case with autonomous systems, which undertake tasks with a degree of independence from their user. As these autonomous systems become more advanced, their capacity for complex tasks also increases, yet with this the opportunity for errors increases too (Parasuraman et al., 2000). As such, while the adoption of autonomous systems is still likely to be informed by many of the factors within TAM, autonomous systems are also particularly reliant upon appropriate trust from their human user, to ensure these systems are used appropriately.

1.3.2 Trust Towards the Designers of Technologies

The line between human-human trust and human-machine trust may also become blurred when we consider that users may also need to consider the reputation of the companies that design or provide these technologies. While trust towards a technology may be based on the usefulness of the service it provides, it can also be based on the reputation of the brand behind the technology (Morgan-Thomas & Veloutsou, 2013), as well as the domain/field in which the technology operates (Pak et al, 2017). For example, Celmer and colleagues (2018) suggest that while the reputation of brands are often discounted in discussions of trust towards autonomous vehicles, the design of these technologies is often a reflection of the brand and their design teams. This builds upon Parasuraman and Riley's (1997) claim that while autonomous systems may reduce the likelihood of errors from humans carrying out a task, this may be replaced with an increased likelihood of errors from the original human designers instead. Culley and Madhavan (2013) also note that while an autonomous system may be inanimate itself, that these systems are still fundamentally the creations of human designers, and as such a reflection of their capabilities, competencies, biases and limitations. Thus, as Celmer and colleagues (2018) suggest, when introducing mass-produced autonomous systems in real-

world settings, any human-automation trust may need to be considered as trust between a human and an automation within the context of the brand that designed the machine.

1.4 Models of Trust Towards Automation

1.4.1 Muir (1987)

Various researchers have put forward models attempting to explain the many factors which inform trust towards autonomous systems. Muir (1987) provided an initial discussion which centred on 4 recommendations for improving calibration of trust towards machines:

1. Improve the human users' ability to perceive the system's trustworthiness

In which the user should be trained to understand how the system works, while the system should be designed to make its decisions more transparent and therefore easier to interpret.

2. Attempt to modify the users' criterion for trustworthiness

In which the boundaries of the systems' performance are clearly defined to the user, by illustrating the system's reliability, to ensure the user has realistic expectations about system competence.

3. Enhance the users' ability to allocate functions and make decisions

In which the balance of power within the human-machine team is addressed, with the human user defined as having significantly greater control and authority when it comes to task-related decision making.

4. Identify and selectively address sources of poor calibration.

In which the inaccurate expectations and beliefs about the system that are held by the human user are addressed within training, in the aim of improving calibration of trust towards the system.

1.4.2 Muir (1994)

Following this up, Muir (1994) expanded on these points when discussing calibrating trust towards machines, by also stating that the initial introductions of autonomous systems should be handled carefully, to ensure that errors are limited which could prevent the formation of mistrust towards the system during critical early stages of use. Additionally, when attempting to recalibrate trust towards systems, Muir (1994) also argues that researchers and managers need to be aware that distrust can be particularly difficult to overcome, especially if a previously trusted system violates the user's trust through task failures.

1.4.3 Lee and See (2004)

The view of trust as a dynamic and malleable variable was central to the model put forward by Lee and See (2004). Within their model, the operator's trust towards the automation, and by extension the way they behave with the system, are dependent on interaction between the operator, the surrounding context and environment, the automation itself, and the interface of the automation. At its core, trust towards the automation, and the subsequent reliance upon the automation form a closed-loop process, wherein interaction and use of the system informs trust towards the system, which then in turn informs further trust calibration. Within this model, contextual factors also play a particularly important role, and can mediate how users' trust informs their reliance on the system. These contextual factors can include the workload and time constraints placed upon the operator, as well as the risks associated with the task outcomes, and the operator's self-confidence. To ensure that the user's trust is appropriately calibrated towards the system, Lee and See (2004) also highlight the importance of information displays within the interface of the automation. As it can sometimes be difficult to directly evaluate the performance of autonomous systems, the information displayed within the interface can help the user to form more appropriate expectations about the system, thereby ensuring more appropriate calibration of trust.

1.4.4 Hoff and Bashir (2015)

Hoff and Bashir's (2015) model also examined the interaction of trust and reliance yet separated trust toward automation into three broad layers. Dispositional Trust relates to stable human-centric factors, such as culture, age, and personality traits, which inform users' general disposition toward technology. This would be reflected in users' general attitudes toward technology, and their propensity to trust new technologies. Situational Trust relates to fluctuating human-centric factors, such as mood and attention, as well as environmental and contextual factors, such as task difficulty, workload, and organizational setting. Importantly, these are all factors which can vary over time. Finally, Learned Trust is split into two separate sublayers: Initial Learned Trust that reflects the user's historical experience of similar systems, and the reputation of the current system, while Dynamic Learned Trust reflects their ongoing experiences of working with the current system. When working with an autonomous system, Learned Trust will likely be informed by the users' ability to interpret the system's decision-making. Additionally, in industrial applications, operators may have previous experiences with other autonomous systems, which may inform their trust toward newly introduced systems. Similar to the model put forward by Lee and See (2004), Hoff and Bashir's (2015) model suggests these three layers of trust combine to inform how users rely upon autonomous systems during collaboration.

1.4.5 de Visser and colleagues (2018)

The view of trust as a dynamic variable that changes over time is also shared by de Visser and colleagues (2018) who illustrate how trust could be repaired after it is lost following poor performance. They argue that too much of the existing literature focusses on the calibration of trust, which focusses on improving the transparency of the systems decision making and communicating the reliability of the system. While they believe this research remains important, their view is that this focus comes from a perception of autonomous systems as are more of a tool that is to be used, rather than an active teammate. Thus, for future autonomous systems which may operate with a greater degree of independence, they argue for more research that examines how trust towards autonomous systems can be repaired.

They propose the use of a transactional model of trust repair, similar to the way trust is understood to be repaired between humans (Tomlinson et al., 2004; Tomlinson & Mryer, 2009). Within this model there are three elements of autonomous system performance, and depending on the outcome, this informs the human user's evaluation of the machine. In the relationship act the machine performs its task, and the outcome is either beneficial or costly to the user's trust. For example, Good Performance is beneficial for trust, whilst Poor Performance is harmful for trust. In the relationship regulation act, a corrective action may be applied to the previous relationship act, which helps to maintain the relationship. These can be immediate or delayed actions, which are either aimed at repairing trust after a costly act, or dampening overly-heightened trust after a beneficial act, in order to ensure optimal equilibrium within the relationship. While dampening trust may seem counterintuitive, this could be beneficial if expectations of the machine are too high, thereby reducing the likelihood of mistrusting the machine. These corrective actions include behaviours such as apologising, providing explanations, and making promises. Finally, in the net victim effect, we can see the influence of these actions and subsequent corrections on the perceptions and experience of the human user, where trust either increases or decreases (de Visser et al., 2018). de Visser and colleagues (2018) also note that individual differences between users will influence the effectiveness of the corrective actions taken by the machine, and stress that these differences should be accounted for when trying to repair the relationship. Thus, within this model, the user's trust towards the autonomous system changes as a dynamic variable, which continuously reflects their perceptions of the system's performance. Ideally, the user's trust will accurately reflect this performance, and the machine may be able to facilitate this through corrective actions, by providing more elaborate insight into system performance.

1.5 Key Factors Influencing Trust Towards Automation

Collectively, it appears that the various factors outlined by these models can be grouped into 3 main categories: Environmental, Mechanical, and Human factors. Within their model of trust, Lee and See (2004) distinguished between the operator, context, the automation and the interface, yet arguably automation and interface can both be collapsed within a greater 'Mechanical' factor given the significant overlap between the two. While these Human, Environmental and Mechanical categories are distinct, they are also capable of overlapping. For example, if a pilot was using autopilot to help navigate difficult weather conditions, trust towards the autopilot would be informed by the performance of the system, yet this performance might be impacted by the weather conditions, thereby illustrating how mechanical and environmental factors can overlap. Equally so, an experienced pilot may be more comfortable in difficult weather conditions, thereby mitigating the influence of the environment, and thereby illustrating the influence of the human factors. For a better understanding, it is worth exploring some of the research that has carried out involving these three categories.

1.5.1 Environmental Factors

Factors associated with the operational environment can affect the task performance of the autonomous system and are often difficult to control. For example, in Selkowitz and colleagues (2017) participants worked with an autonomous squad member robot, which provided them with feedback on environmental elements, such as weather and terrain. While the experiment was simulated, participants trust towards the robot was highest when provided with a combination of different cues, including feedback on environmental factors such as hazards, as well as cues on system resources such as fuel, and explanations of system decision making, such as the intentions and motivations (Selkowitz et al., 2017). Thus, participants trust towards the system was, at least partially, informed by their understanding of environmental factors when evaluating system performance. In Hoff and Bashir's (2015) model, these environmental factors fit within their larger framework of Situational Trust, in which trust towards the system is partly informed by contextual factors within the environment. However, Hoff and Bashir's (2015) Situational Trust also includes transient factors that are intrinsic to the human operator, such as mood, attentional capacity, and self-confidence. While these are all factors that can change with time, grouping them together like this blurs the line between human and environmental factors, and therefore for the purposes of this review we will consider human and environmental factors separately.

1.5.1.1 Task Difficulty and Risk

While environmental factors can impact upon the performance of the autonomous system, they may also change how the operator uses the autonomous system. In a study involving an autonomous letter detection aid, Schwark and colleagues (2010) report that users were more likely to rely upon the system in trials where success was framed as being more important. Moreover, they were also more likely to use the system in trials that were described as being more difficult to complete (Schwark et al., 2010). Likewise, in a study involving a military convoy leader task, when the convoy was in situations where it was most vulnerable to attack, operators were reportedly more likely to use the guidance of an autonomous aid, rather than the guidance of a human aid (Lyons & Stokes, 2012). This would suggest that when faced with higher-stakes tasks, some users may become more reliant upon automation.

However, other research suggests that these environmental factors can also dissuade some users from trusting and relying upon autonomous systems. When using a GPS-based route planning system, operators were reportedly less likely to trust the system's suggestions when faced with more serious hazards, such as burning buildings and riots, compared to when facing lower stakes risks such as traffic jams (Perkins et al., 2010). Similar findings were reported from a study which explored risk through financial incentives. Satterfield and colleagues (2017) examined how perceived risk informed operators' willingness to trust an autonomous UAV system during a UAV management task. Risk was introduced to their laboratory setting across low and high-risk conditions, by using a points-based system wherein participants were penalised for task failures and rewarded for task successes, which directly contributed to their final financial reward. In the high-risk condition where there were greater financial deductions associated with failures, participants were more likely to intervene in the UAV's area of responsibility, suggesting that they had less faith in the system completing its task successfully (Satterfield et al., 2017). This illustrates how environmental factors such as task difficulty and risk can inform the ways in which operators may decide to trust and use autonomous systems. However, this may be further complicated by the user's ability to evaluate the performance of the system.

When faced with complex tasks, it may be difficult to accurately evaluate the performance of an autonomous system. This idea of task ambiguity was explored by Merritt and colleagues (2013) in a study involving trust towards an autonomous baggage scanner. Participants collaborated with the scanner to detect weapons within images of luggage, which had contents that were either empty or cluttered, meaning that weapons were only easy to identify in some of the trials. By doing so, this gave the researchers an

insight into how users placed trust in the system when the performance was not easy for them to evaluate. Merritt and colleagues (2013) report that when the performance of the scanner was ambiguous, trust towards the scanner was higher in individuals with a higher self-reported propensity to trust machines. This suggests that some individuals may have a bias towards technology (discussed in more detail later) and are subsequently more favourable towards autonomous systems even when their performance is difficult to evaluate. This also illustrates how environmental factors can interact with human-centric factors to inform users' trust towards autonomous systems. Unsurprisingly, when system performance is ambiguous, it can also lead to more errors from the human-machine team. In a similar study, Chavaillaz and colleagues (2020) explored trust towards an automated baggage scanner that provided false alarms which were either 'plausible' or implausible'. Participants were more reportedly likely to accept false alarms from the aid when it provided alerts for items that were plausibly similar to items that were on a prohibited list. In contrast, this was not the case for implausible false alarms, where the cued luggage item had little-to-no resemblance to items on the prohibited list. This illustrates how users' reliance upon automated systems can become complicated when system performance is difficult to evaluate. Arguably, manipulations which introduce ambiguity into system performance should be more prevalent within the wider trust in automation literature, as they help simulate conditions which could be typical within real-world applications of autonomous systems.

1.5.1.2 Operator's Workload

While environmental factors such as task difficulty can inform trust towards autonomous systems, it is also worth considering the workload of the operators themselves as an environmental factor. Given autonomous systems have many industrial applications, it makes sense to consider that the people who will be using these systems will be doing so within the workplace, where they may have competing priorities and tasks. When working with machines, operator workload is typically measured subjectively, using questionnaires such as the NASA Task Load Index (NASA-TLX) (Hart & Staveland, 1988). The NASA-TLX captures the physical, mental and temporal demands placed on the user, their subjective evaluation of their task performance, and the amount of effort and frustration they experienced in the task (Hart & Staveland, 1988). The correct use of autonomous systems has the potential to significantly improve the workload of the user, as the autonomous system can perform more repetitive tasks and allows the operator to focus on other, more complex tasks. For example, McKinney and colleagues (2020) suggest that their image classifier, which is trained to detect breast cancer, could lower oncologists' workload by up to 88%. Such benefits could be seen in the findings of Lyle and colleagues

(2018), who report that the use of CDSS can help to reduce the workload of healthcare practitioners, particularly when faced with complex cases requiring greater attention to detail. However, these systems will only be beneficial to their user if they can perform their tasks correctly.

When an autonomous system has low reliability, the operator may need to work harder in order to compensate for the machine's mistakes, which can increase the user's mental workload (Chavaillaz et al., 2016). This reliability itself is likely to be affected by the operational environment of the human-machine team, where other environmental factors such as extreme weather may reduce system reliability. Biros and colleagues (2004) suggest that the perceived predictability of autonomous systems, and the user's dependability on the system can directly inform trust towards the system. If a system is unpredictable, the user may have less trust in the system, and therefore may be less likely to use the system (Hoff and Bashir, 2015). However, Biros and colleagues (2004) also suggest that when a user is given a heavier workload, they may be more likely to use an autonomous system, even if they do not trust the system. This could be particularly problematic if it leads to the user accepting incorrect decisions from the system, particularly given some autonomous systems are employed in high-stakes settings. Thus, the workload of the user should be considered as an environmental factor which can significantly influence how operators trust and use autonomous systems. While autonomous systems can help to reduce their operators' workloads, these benefits may only occur when they are used correctly.

Collectively, in line with Hoff and Bashir's (2015) framework of Situational Trust, the existing research suggests that environmental factors can influence users trust towards automation and can also inform the ways in which they use and rely upon the system. In many ways, these environmental factors will also contribute to the design of the system itself, within which there are a multitude of factors that can further influence trust.

1.5.2 Mechanical Factors

1.5.2.1 Performance and Reliability

Amongst mechanical factors, the central and most important factor is the performance of the autonomous system (Muir, 1987; Parasuraman & Riley, 1997; Lee & See, 2004; Hoff & Bashir, 2015). At the most fundamental level, system performance reflects the system's ability to carry out the tasks it was designed to undertake (Hoff and Bashir, 2015). When viewed over time, system performance becomes system reliability, which is an estimate of how well the system repeatedly performs its task and reflects how predictable the machine will be at any given time (Biros et al., 2004; Chavaillaz & Sauer,

2017; Chavaillaz et al., 2016; Sauer & Chavaillaz 2017; Hussein et al., 2019). As such, users' trust towards autonomous systems is closely related to their perceptions of the system's accuracy when it is carrying out tasks (Berkeley et al., 2015; Yu et al., 2019; Zhang et al., 2020).

Given trust is a dynamic concept, human users will typically update their trust in response to changes in system performance, lowering their trust when performance is poor, and elevating their trust when performance is good (Desai et al., 2013; De Visser et al., 2018). However, if poor performance is too prevalent, or if the errors have critical consequences, then this may culminate in disuse of the system entirely (Parasuraman and Riley, 1997; Berkeley et al., 2015). This relationship is further complicated by factors such as overreliance, in which an operator may distrust a system, yet is unable to undertake the system's responsibilities, in which case they may rely on the system, but do not trust it (Chavaillaz et al, 2016; Biros et al., 2004; Hussein et al., 2019). When this occurs, the operator may second guess a system, even if it is sometimes correct, leading to more errors from the human-machine team, ultimately at the potential detriment to the cognitive capabilities of the human operator (Goddard et al., 2014; Chavaillaz et al, 2016). At the same time, if the operator does not use the system correctly this can in turn affect the reliability of the system (Ozdemir & Kumral, 2019). Ultimately, the challenge for trust in automation research is to ensure that human users can calibrate their trust towards the system, so that it accurately reflects the system's performance capabilities. As suggested by both Lee and See (2004) and Hoff and Bashir (2015), the performance of the autonomous system can be conveyed more clearly through different features displayed within the system's interface, which can help the user to better understand its decisions. By providing the user with more detailed information about system performance, we are increasing the transparency of the system's decision making.

1.5.2.2 Transparency and Explainability

Whilst autonomous systems are becoming more capable of undertaking complex tasks, they are often criticised for being uninterpretable 'black box' systems, particularly when neural networks are involved (Abdul et al, 2018; Ribera & Lapedriza, 2019). As such, there has been increasing interest in promoting the explainability of autonomous systems, in order to make their decision making more transparent. Explainability is characterised as the system's ability to convey the reasoning behind its decision-making, thereby facilitating greater understanding from human users (Gilpin et al., 2019). In theory, if a human user can better understand the decisions of an autonomous system, they should be able to trust the system more accurately. There are a variety of ways in which system decision making can be conveyed to the user, however the most commonly used method

across many different autonomous systems appears to be System Confidence Information (SCI).

SCI is a form of decision support information, which reflects an autonomous system's confidence in carrying out its task, often as a predicted probability of the decision being correct (Zhang et al., 2020). The presence of SCI within the interface of autonomous systems has previously been found to inform users' evaluations when working with autonomous systems. For example, when SCI was presented in the interface of an autonomous handwriting reader, users were more likely to accept the decisions of the system (Verame et al., 2016). Similarly, in a study involving an autonomous robot navigating environments of varying difficulty, participants provided with SCI were more likely to assist the robot than participants without (Desai et al., 2013). Moreover, the same participants were better at dynamically modulating their trust towards the robot in response to cues of high, medium, and low system confidence, with trust decreasing when confidence was low, and increasing when confidence was high (Desai et al, 2013). This suggests that SCI could help users to better calibrate their trust towards autonomous systems and may even help with recovery of trust following poor system performance.

While SCI appears to inform how users perceive the performance of autonomous systems, it also appears to improve the overall performance of human-machine teams. Pilots operating a flight simulator performed less errors and had fewer stalls, when their decision support system provided SCI that reflected the system's recent performance (dynamic), compared to when it only provided an overall average summary (static) (McGuirl and Sarter, 2006). This suggests that operators incorporate SCI when working with autonomous systems, which may ultimately improve human-machine team performance. However, McGuirl and Sarter (2006), also note that across both static and dynamic feedback conditions, when system confidence was lower, participants tended to reject the advice of the system, and instead made the opposite decision. The authors suggest that their participants misinterpreted low confidence as lower performance, which is problematic, given low confidence does not automatically coincide with low/poor performance. A similar finding was also reported by Verame and colleagues (2016), where participants were more likely to accept the decisions of the autonomous document reader when confidence was described as "very high", compared to when described as "medium", "low", and "very low" confidence. Therefore, when working with autonomous systems, human operators' evaluations appear to be influenced by the provision of SCI. However, it seems that this information should be presented in a meaningful way, so that it illustrates decision making without undermining perceptions of system performance.

1.5.2.3 Information Complexity

When collaborating with an autonomous system the use of decision support information, such as SCI, may improve trust towards the system, and lead to a more productive human-machine team. However, there are a variety of ways in which SCI could be conveyed within the system's interface. If a cue of system confidence is too simplistic, users may not appreciate it, yet if it is too complex, users may feel overwhelmed and disuse the information. Theoretically, an optimal cue of SCI would provide the operator with enough information to clearly illustrate the performance of the system yet would do so without being overly complex. Providing the user with too much information may unnecessarily increase the user's workload, thereby mitigating the benefits of using the automation.

Evidence suggests that when working with some autonomous systems, human users may prefer to receive decision support information in simpler, less complex formats. Koo and colleagues (2015) examined the influence of multiple levels of information complexity on user trust towards an autonomous driving aid. Whilst operating a driving simulator, drivers were given messages that explained: *how* the aid would behave ("it will make a left turn"); *why* the aid would behave ("there is an obstacle"); or *how and why* the machine will behave ("it will make a left turn because there is an obstacle"). When the aid performed unsafe driving behaviours, drivers reportedly preferred the messages that only explained *why* the aid behaved as it did, which coincided with the lowest reported anxiety and the highest reported trust towards the system. Contrary to their hypotheses, combined messages of *how* and *why* the aid was behaving reportedly made users more anxious, which the authors speculate to come about from the result of information overload.

Likewise, information complexity was also explored in Selkowitz and colleagues' (2017) study, where participants monitored an autonomous robotic squad member while it navigated different environments. They used multiple different interfaces, which incrementally increased in the complexity of information provided to users. Their participants were reportedly more likely to trust the squad member when they used an interface which displayed the most detailed situational information, such as the system's motivations and predicted task outcomes. However, their trust was not increased when this information was also augmented with a cue illustrating the degree to which the system was uncertain about this information (Selkowitz et al., 2017). This suggests that adding more complex information about system performance was beneficial for user evaluations up to a certain extent, after which excessive information may not improve trust, similar to the findings of Koo and Colleagues (2015). Thus, decision support information appears to

influence trust towards autonomous systems when presented in different ways, yet higher complexity does not appear to automatically foster higher trust.

Rudin (2019) suggests that when designing autonomous systems, there should be a distinction made between systems that are designed to be explainable and those that are made to be interpretable. Much of the previous research in trust towards automation appears to involve making systems explainable, wherein the decision making of the machine is retroactively explained to the user, through cues such as SCI. However, Rudin (2019) argues that developers and designers should put more emphasis on making systems that are inherently interpretable, so that users can more easily understand them without explanations, particularly when used in high-stakes settings. Such systems would require sophisticated interfaces designed to provide their users with optimal transparency yet appear to be something that may only become more prevalent in future research.

While users may benefit from working with autonomous systems that provide greater transparency, the influence of increased transparency on trust may only go so far. Indeed, Papenmeier and colleagues (2019) report that users' trust towards an automated text classifier appeared to be based more on the system's accuracy, rather than the system's explanations for its decisions. Similarly, Wright and colleagues (2019) report that users' trust towards an autonomous squad member was more profoundly informed by the system's reliability, and much less impacted by the transparency of the decisions made by system. Unsurprisingly, these mechanical factors may be limited in their ability to compensate for losses in trust if the overall performance of the system is poor. Ultimately, while system performance is central to trust towards automation, it is important to consider that factors related to the human operators are also a significant contributor for trust towards autonomous systems.

1.5.3 Human Factors

For all the other environmental and mechanical factors that can inform trust towards automation, some of the variance in operators' trust will also be explained by individual differences amongst different human users. Hoff and Bashir (2015), Lee and See (2004), and de Visser and colleagues (2018) all recognise the importance of human factors in shaping trust towards automation. Hoff and Bashir (2015) characterise most human factors within the Dispositional Trust component of their model, which accommodates factors such as age, culture and personality traits. Likewise, Lee and See (2004) also consider self-confidence, cultural differences, and predisposition towards trusting as human factors which can inform trust towards automation. These human factors can shape how operators use and evaluate autonomous systems. For example, novice operators

benefited more from using an automated aid when using an X-ray scanner to process cabin baggage, in comparison to expert operators (Chavaillaz et al., 2019). Moreover, for the same expert operators, their elevated detection rates within the task were not reportedly influenced by the use of the automated aid, which illustrates the importance of experience and skill levels between human users (Chavaillaz et al., 2019). Amongst operators, shortages in experience and skill are something that can often be addressed through training and prolonged exposure to the task. However, it may be more difficult to address biases towards technology, which are one of the most common human factors within the trust in automation literature (Challen et al., 2019; Rice et al., 2017; Lyell & Coiera, 2017).

1.5.3.1 Biases Towards and Against Technology

Automation Bias is the belief that automation performance is inherently superior to human performance, where users may tend to overrule their decisions in favour of those made by the system, on the basis that a machine is more likely to be correct (Cummings, 2017; Goddard et al., 2012; Lyell & Coiera, 2017). While expert operators may be expected to be less susceptible to Automation Bias, evidence suggests they may be just as vulnerable as novices, and in some cases may be even more likely to rely on automation (Mosier et al., 2017). Automation Bias represents a particular problem in high-stakes settings, such as within the healthcare industry, where automated Clinical Decision Support Systems help healthcare practitioners to make diagnoses (Challen et al., 2019; Sujan et al., 2019; Goddard et al., 2014; Lyell et al., 2018). In a study involving a simulated clinical decision support system with a reliability rate of 70%, healthcare practitioners with Automation Bias overruled correct diagnosis answers provided by the system in 5.2% of the cases (Goddard et al., 2014). While this number may seem small, in a real-world scenario, this bias could result in a patient being misdiagnosed by the health practitioner, who rejects a correct decision from the system out of mistrust.

Similar to Automation Bias, the Perfect Automation Schema (PAS), proposed by Dzindolet and colleagues (2002) suggests that human users are vulnerable to forming the belief that automation performance is almost perfectly reliable. Individuals who score high for PAS are more likely to trust autonomous systems but are also more likely to have heightened expectations for the system's performance (Lyons et al., 2019). If this schema is violated, operators may stop using the autonomous system and instead become more self-reliant in the task (Dzindolet et al., 2002). Support for this was reported by Merritt and colleagues (2015) who found that individuals who scored higher on criteria for PAS displayed more intense decreases in trust after witnessing automation errors. Inversely, positive biases towards machines can also inform how users interpret autonomous system

performance, even if the performance of the autonomous system is difficult to evaluate. As discussed previously, individuals with higher self-reported scores of propensity to trust machines were more likely to trust the ambiguous performance of the autonomous baggage scanner used in Merritt and colleagues' (2013) study. Collectively, this illustrates how perceptions of autonomous system performance, and by extension willingness to use and rely upon automation, can be partly informed by individual differences between operators. Interestingly, while biases in favour of technology can improve trust, research suggests that trust towards automation can also be improved by the presence of human-like anthropomorphic traits.

1.5.3.2 Anthropomorphism

Anthropomorphism is the introduction and application of human traits and behaviours to non-human entities, and is commonly used in children's cartoons and advertisements, and even in more abstract concepts such as theology (Duffy, 2003; Epley et al., 2007). This preference for human-like traits and behaviours has historically been used by humans to rationalise the behaviours of non-humans, as a way to better understand things beyond our control, such as wild animals and even weather systems (Mitchell & Thompson, 1997; Epley et al., 2007). As smarter technology has become more ubiquitous, designers have also introduced anthropomorphism as a means to help users accept these technologies, particularly with systems such as social robots (Duffy, 2003).

While these traits, such as genders and names, may not necessarily benefit mechanical performance, some evidence suggests that they may improve user reliance and trust towards the system. In a study that examined attitudes towards different computer agents, participants showed a preference for agents with more anthropomorphised features (de Visser et al 2012). When asked about their perceptions of computer agents that varied along an anthropomorphic spectrum, agents with more human-like avatars were perceived as more knowledgeable than agents with computerised avatars (de Visser et al 2012). Moreover, when the reliability of the agent was low, computerised avatars had larger drops in trust compared with more human-like avatars (de Visser et al., 2012). Similarly, Waytz and colleagues (2014) report that individuals who interacted with different autonomous vehicles preferred ones with more anthropomorphic features, such as names, voices and genders, as opposed to vehicles without. Even when a system does not include physical or visual anthropomorphic features, the inclusion of behaviours that mimic the behaviours of other humans is also reported to shape trust in human-machine interactions. When interacting with a computer system that behaved with good etiquette (i.e. not interrupting the user when they were busy) human-machine team performance was reportedly improved, even when the system's reliability was low (Parasuraman and Miller, 2004).

This was not found to be the case when operators interacted with a machine designed with poor etiquette, such as interrupting the user during tasks (Parasuraman and Miller, 2004). This suggests that when designing autonomous system interfaces, the adoption of traits and behaviours that mimic those of other humans may benefit human-machine team performance.

Further evidence for anthropomorphic traits may even indicate that there is a biological basis for humans' preference for human-like features. de Visser and colleagues (2017) report that when participants were administered oxytocin, a hormone that is released through contact with other humans, there was an increase in trust and compliance with an anthropomorphised avatar. This also led to better collaborative performances between the human users and the automated agent in a shared task de (Visser and colleagues, 2017). Interestingly, this was not found to be the case when the agent was represented with a computerised avatar that did not have human-like features, suggesting that may be an interaction between oxytocin and anthropomorphised features, which may influence trust and cooperation. While evidence suggests that some human users may prefer autonomous systems which are more anthropomorphic, ultimately the use of human-like features may only be appropriate in certain contexts or tasks. At the end of the day, most autonomous systems are employed to undertake a series of complicated tasks, and anthropomorphic features may distract or even hinder team performance in some settings.

1.5.3.3 Perceptions of Autonomous Systems as Teammates

On a more general level, trust towards automation may also be influenced by the way that people interpret the role of the autonomous system within human-machine teams. When considering how technology is used in different human–machine teams, Larson and DeChurch (2020) make a distinction between technology and agents. Technology is something that is used by teams to achieve their goals, much like a tool, while agents fill a distinct role within the team which goes beyond mere augmentation, and can inherently improve the team's performance as a result (Larson & DeChurch, 2020). For agents, they also draw a distinction between robots, which are agents with embodied physical characteristics, and AI which are disembodied agents that perform tasks that traditionally require human intelligence, such as visual identification and decision-making (Larson & DeChurch, 2020). Thus, when working with autonomous systems human users may use similar distinctions between different types of technology, when attributing responsibility and blame.

In a recent study, when participants read descriptions of human-robot task failures, humans tended to be rated as most blameworthy, followed by robots, with environmental factors rated least blameworthy overall (Furlough et al., 2019). However, in scenarios that

distinguished the robot as being autonomous, the robot was more likely to receive a higher proportion of blame for task failures, compared to when described as non-autonomous (Furlough et al., 2019). This illustrates how human beliefs can influence trust, and suggests users' evaluations are partly based on the perceived capacity of the autonomous system. This can become even more complex, if we consider that these human factors can also interact with environmental factors. When seeking advice, the type of task being undertaken, and the perceived capacity of the adviser can inform how individuals engage with different types of advisers. In a study examining the role of agent and task type, when individuals were presented with emotion-based social tasks, they were more likely to assume human agents had more expertise, and therefore sought their advice more frequently (Hertz and Wiese, 2019). However, when faced with a number-based task, the individuals were instead more likely to seek the advice of robot and computer agents, indicating they believed robots were better suited to more analytical tasks (Hertz and Wiese, 2019). Thus, the users' perceptions of the capabilities of the autonomous system, and their beliefs about the system's role within the team may also help shape their trust towards the system.

1.6 Conclusion

The purpose of this literature review was to provide an overview of the existing trust-in-automation literature, explore the dominant theoretical models of trust towards automation, and identify the key factors which could be explored within a series of studies in this thesis. As this research is cross-sectional, sitting between Psychology, Computing Science, and Engineering, there were many areas of literature that were overlapping. To keep this literature review concise, I focussed on literature that has engaged with popular models of trust in automation, such as those put forward by Lee and See (2004) and Hoff and Bashir (2015). This literature review may have been improved by including more research from the autonomous vehicle literature, as well as including more research involving social robotics. However, both of these fields are already well established and incredibly diverse, and providing a detailed summary of both fields would have increased the length of this literature review significantly. Ultimately, I believe that the literature I did include was more than enough to shape and justify the following studies in Chapters 2-4. In summary trust towards autonomous systems is an area of growing interest, which has increased in line with recent advances that have seen autonomous technologies become both more intelligent, and more pervasive within modern workplaces. Multiple models have been put forward attempting to map out the various ways that a user's trust towards an autonomous system can be influenced. At the broadest level, these factors can be

grouped into three main domains of influence: Human factors relating to the user; Mechanical factors relating to the autonomous system; and Environmental factors relating to the contextual operating environment of the human-machine team.

Chapter 2: Exploring the Influence of Trust Towards Digital Services on Users Willingness to Share Information

2.1 Abstract

Digital economies are populated by services such as ecommerce, streaming platforms, and social media services. These digital services have reached widespread popularity in many developed nations, and play an important role within modern life. Digital economies rely upon users' willingness to share their information, which is influenced by trust towards these digital services. Social media services in particular are dependent on their ability to collect and analyse users' data in order to provide engaging content, whilst simultaneously informing advertisers, vendors, and other digital services about consumer interests. High-profile, controversial events throughout the 2010s have raised questions about the data collection practices of digital services, and could damage users' trust towards sharing their personal information and data. We present qualitative data from a project which examined the ways individuals characterize their trust towards social media and other digital services. Thematic analysis of participants' responses illustrated participants trust towards digital services through two main themes: *Information Security* and *Service Transparency. Information Security* highlighted participants' considerations for how secure their information was when using the service, and *Service Transparency* highlighted participants' beliefs about how these services use their information, and the intentions of the company providing the service.

2.2 Introduction

With increased internet accessibility at the turn of the century, Social Media Services (SMS) such as Facebook, Instagram, Twitter and YouTube saw unprecedented levels of engagement from users. Estimates suggest that this rate of uptake for SMS far outstrips other historically important technologies, such as the automobile and the television (Desjardins, 2018). While most SMS are free to use, they rely on access to users' information and data as a core feature of their business models. As SMS users engage and connect with each other, their behavioural data is collected and analysed to determine the types of content they are interested in, allowing these services to provide similar content to keep users engaged. This data is also provided to external advertisers to help them understand the services and goods that may be of interest to these users (Lipsman et al, 2012). Through this, SMS provide a cornerstone of the digital economy, by

introducing consumers to vendors and their products by advertisements structured around their interests (Alhabash et al., 2017; Lee et al., 2018). Within the wider digital economy, other digital services, such as e-commerce, media streaming services, and online banking have all seen similarly explosive increases in usage (Wu et al., 2017; Adhikari et al., 2014; Özlen & Djedovic, 2017).

A great deal of research has already examined how users' wellbeing is influenced by the use of digital services, and in particular SMS, with some evidence suggesting these services improve users' self-esteem and help users to maintain friendship networks (Hampton et al., 2011; Mazurek, 2013; Park et al, 2013). Conversely, other evidence suggests the use of SMS can be potentially detrimental to wellbeing, particularly with users who become over-reliant and compulsively use these services (Blackwell et al., 2017; Hawi & Samaha, 2017; Woods & Scott, 2016). Moreover, the direct communication capabilities of SMS also facilitate cyberbullying, with extensive research examining the causes and effects of these behaviours amongst users (Parris et al., 2020; Machackova et al, 2013; Sabella et al., 2013). In the wake of high-profile data breach controversies (Cadwalladr, & Graham-Harrison, 2018; Lewis, 2014; BBC News, 2017; BBC News, 2013), there has also been increased interest in the trustworthiness of companies that provide SMS and other digital services. These data breach controversies have publicly illuminated how users' data can be misused by the companies that provide these platforms. Thus, this study sought to use qualitative research to explore how individuals characterize their trust towards digital services, with a particular focus on SMS given their popularity. It also sought to understand how this trust then informs their willingness to share their personal information when using these services.

2.2.1 The Importance of Trust for Digital Services

Trust is a dynamic and complex variable informed by a variety of factors, experiences and events (Hoff and Bashir, 2015; Lee and See, 2004; McKnight et al., 2002). Digital ecosystems require continuous access to users' information and data, yet if users do not trust the companies involved, they will be less likely to use and engage with their services. There are a variety of factors underpinning trust towards digital services, which can stem from the consumer, the vendor and the vendor's website (Chen And Dhillon, 2003). An example of the importance of this trust can be seen with eBay, the digital marketplace that facilitates consumer-consumer sales. Whilst eBay provides a secure platform for these transactions to occur, the rating system for individual users' reputations allows buyers and sellers to evaluate prospective transactions, and decide whether or not to

trust the other user, helping to form what is described as a ‘Community of Trust’ (Boyd, 2002). In this environment trust is crucial, yet if digital services like eBay are unable to retain the trust and engagement of users, they may cease to survive (Boyd, 2002).

Trust towards digital services is not only required for services that are focussed on e-commerce. When users share information about themselves, such as commenting on public websites, trust in digital services also predicts a greater likelihood of disclosing personally identifiable information, such as names and locations (Gustavo and Mesch, 2012). Similarly, on SMS users are more likely to share and disclose personal information if they are more aware of how their information is used by the SMS provider (Benson and colleagues, 2015; Tufekci, 2015). This suggests that trust can inform the way that people use digital services; if a user has more trust in the service, they may be more likely to share their personal information. However, recent high-profile stories within the news about SMS providers have brought attention towards the way our data can be used (and misused) by the digital service providers.

2.2.2 The Rise of Public Distrust Towards Social Media Services

While SMS are intended to allow users to maintain social connections with other users, they can also facilitate the spread of false and misleading information. SMS providers have been criticised for creating filter bubbles, otherwise known as ‘echo-chambers’, in which specific, extreme, and often disingenuous views and ideas can be propagated amongst subgroups of users (Pariser, 2011). These echo chambers are believed to have helped spread many false articles and narratives about political and international entities, known collectively as ‘Fake News’ (Allcott & Gentzkow, 2017). The nature of these echo chambers first received widespread attention in the wake of the Cambridge Analytica Scandal, in which a third-party organization (Cambridge Analytica) gained unauthorised access to large amounts of Facebook users’ data (Cadwalladr, & Graham-Harrison, 2018). While the larger extent and implications of this are still relatively unknown, Cambridge Analytica are believed to have potentially used this data to influence the opinions of voters prior to important elections and political referenda, through misleading advertisements and articles (Cadwalladr, & Graham-Harrison, 2018). More recently, the ‘Qanon’ conspiracy theory has propagated online through SMS, spreading many diverging and extreme right-wing beliefs (Amarasingam, & Argentino, 2020). Alongside other extremist factions, Qanon followers were implicated in the 2021 insurrection of the U.S. Capitol building (Amarasingam, & Argentino, 2020; Dalsheim & Starrett, 2021). While the investigation into the events of the 2021 insurrection are still ongoing, SMS platforms such as Facebook and Twitter have been criticised for not

intervening more decisively when users used their platforms to spread false information and coordinate these attacks (Sung and Klein, 2021). A similar example of this can be seen in the way that misinformation has spread during the COVID-19 pandemic, with efforts to combat the virus undermined by SMS users' sharing inaccurate beliefs about vaccines and epidemiology (Limaye et al., 2020). Similar relationships between SMS and misinformation were also reported during the 2014 Ebola epidemic (Oyeyemi et al., 2014), suggesting this is not exactly a new problem. As a result of these events, there is now increased interest in how people share and consume information online, and whether the companies that provide these platforms should regulate their content more rigorously. Ultimately, these questions also raise questions about the trustworthiness of the companies that these platforms.

2.2.3 Characterising Trust in Digital Services

The relationship between trust and digital services appears to be complex. Individual differences between different users will influence how likely they are to place trust in digital services, whilst features in the design and functionality of the website/app will also shape the trustworthiness of the service. (Chen And Dhillon, 2003). For SMS, this characterisation of trust may also extend to include their trust towards other users, and may also vary between different SMS platforms, where they may have different social networks (Warner-Søderholm et al., 2018). Trust towards SMS may also be informed by the content that users consume when using these services, particularly when seeing misinformation. For example, individuals who reported higher trust in the news they receive from social media were also more likely to believe long-standing conspiracy theories (Xiao et al., 2021). However, other evidence suggests that the spread of misinformation might be linked more to users' trust of other users, rather than their trust towards the platform itself. A recent study by Sterrett and colleagues (2019), suggests that SMS users' trust towards news articles may be more likely to be influenced by the trustworthiness of the individual who shares the article, rather than the trustworthiness of the source of the article itself. This echoes similar research by Anspach and colleagues (2020), who report that when interacting with news content on SMS, users typically have higher trust in the journalists who produced the news, rather than the individuals who shared the news. This would suggest that users may be more critical of the person who shares an article, rather than those who actually write and publish the article itself. Collectively, trust towards digital services can be thought of in a number of different ways, and different users may find particular factors more significant when characterising of their trust.

While there are questions about how much trust users should place in SMS, individuals continue to use them, with some more popular than ever. Despite Twitter facing considerable criticism over its recent handling of misinformation (Hiar, 2021), Twitter has also reported continued growth in their active users (Kastrenakes, 2021), suggesting that users' trust towards SMS may not be broken by scandals and controversy. Likewise, recent examples of data breaches in digital services like iCloud, Equifax, and Adobe have highlighted how vulnerable millions of users' data actually is when using other digital services (Lewis, 2014; BBC News, 2017; BBC News, 2013). Nonetheless, given the important role that SMS and other digital services play in modern economies (Alhabash et al., 2017; Lee et al., 2018), it is important to understand how trust influences usage, engagement and attitudes towards sharing information with these services. Given this, we conducted a qualitative study which used focus groups to better understand the question: **RQ1:** *How Do People Characterize Trust towards SMS and Other Digital Services?*

2.3 Methods

2.3.1 Design

We wanted to explore trust towards SMS and other digital technologies using qualitative methods as a way of letting participants share their experiences and views on a topic that has been traditionally measured through quantitative methods. We considered using diaries, which participants could have filled in with their views over an extended period of time. We also considered using individual semi-structured interviews to give each participant a chance to share their views. Ultimately, focus groups were chosen as a way of getting participants to interact with each other, which we believed would help foster debate and discussion. Participants were asked a series of semi-structured questions in a focus group setting, during which they were free to deviate from questions and expand upon their answers. These focus groups were conducted during the initial stages of a larger project, in which our collaborators (Qumodo) were developing an artificial intelligence tool intended to help users retrieve and remove their explicit images that have been inappropriately shared online. Here, we focussed primarily on understanding how users characterize their trust towards SMS and other digital services, and how this informs their willingness to share their information with them.

2.3.2 Participants

16 participants (10 female) were recruited to take part in this study (ages 23-34, average=30), and focus groups were small in size (3-4 members). Participants were offered

compensation at a rate of £6ph. Participants were a mix of students and professionals from a variety of working backgrounds. Professionals were recruited externally through our external collaborators, and were from backgrounds in finance, technology research, marketing, education, and clinical psychology.

2.3.3 Interviews

All focus groups were conducted between July and October 2018. Participants were asked a series of questions in a semi-structured interview format, relating to: 1.) their use and engagement with SMS and digital services, 2.) their attitudes towards sharing information online, and 3.) their trust and views towards technology, particularly Artificial Intelligence. The initial questions for each topic were intended to be as broad as possible, in order to not lead participants' views in a particular direction, with more specific questions asked towards the end of each topic. For the full list of the questions used in these focus groups, see Section 2.6. As they discussed these topics, participants were free to expand and diverge on issues they felt were important. Participants were debriefed following completion of focus groups and reminded of their rights to anonymity and their control over their data. All focus groups were recorded on a dedicated digital voice recorder, which were then transcribed for analysis. All participants were provided with pseudonyms to ensure their anonymity when sharing their views and experiences.

2.3.4 Procedure

Interested participants were then invited to attend focus groups, which were either held at the University of Glasgow's School of Psychology, or our collaborator's office in London (Qumodo). In a semi-structured interview format, participants were asked questions regarding their general attitudes towards social media and digital services, the ways in which they engage and use these services, and their perceptions and experiences of information sharing and security. Throughout this, participants were free to divulge as much or as little they deem appropriate and were free to expand on issues that they felt were important. Participants were given debriefing forms following completion of the focus groups and were again reminded of their rights to anonymity and to access/remove their data. Following the focus groups, the recordings were transcribed for qualitative analysis. Thematic analysis was then used to identify themes in the collective views of the participants.

2.3.5 Data Analysis

To understand the role of trust in how users engage and share information when using social media and digital services, we conducted thematic analysis on transcripts of

participants answers, in line with Braun and Clarke (2006). Three members of the research team worked independently to open-code the data from an inductive analysis perspective that was independent of an existing theoretical framework. Codes were generated using meaningful statements taken at the semantic level throughout the transcript, taking the meaning of participants' views at surface level. These codes were then grouped and developed into subthemes that represented participants' collective responses and experiences. The researchers then met to review their independent analysis, with discussion on how these subthemes could be combined to represent similar groupings of thought, and which larger themes could be created to reflect participants' views and experiences more cohesively.

2.4 Findings

We posed the question: **RQ1:** *How Do People Characterize Trust towards SMS and Other Digital Services?* Thematic analysis of participants' responses allowed for characterization of trust towards SMS and digital services into the 2 themes of: *Information Security* and *Service Transparency*.

2.4.1 Theme 1: Information Security

In the theme of Information Security, when sharing personal information and data with SMS and digital services, participants often considered the security of the platforms they were using, as well as the control they had over the information they shared. Unsurprisingly, a significant subtheme of this was *Platform Security*, which covered participants' evaluations of how safe their information is with these services. In a similar subtheme, *Other Users*, participants also discussed the role that other service users may play in information security on SMS. When sharing content with friends and other users, these individuals may share the content further, thereby spreading it to a wider audience. Lastly, the other subtheme within Information Security was *Lack of Control*, which captured some participants' views when they felt coerced into using these services, as well as their strategies for using services they distrust.

2.4.1.1 Platform Security

For many participants, their trust towards SMS and other digital services was based, at least partly, on the perceived security of the service. Simply put, participants expressed distrust towards services that they deemed insecure. Often, this appeared to be something that became apparent to them over time.

Liam: *'I think there's like […] there's an inherent level of trust that's… starting to erode as we realise the consequences of… and the lack of security that exists.'*

A common trend that participants reported was the regulation of information they shared with services they believed were insecure. This regulation of information sharing was prominent throughout the subthemes within *Information Security*. If participants felt that they couldn't trust an SMS or digital service, they were much less inclined to share their information with it.

Ken: *'Its about whether they meet the expectations that I have for them, Facebook is a really good example, when I first joined Facebook I had everything on there, and then I learnt in time its actually not safe, and […] I took some stuff off there, I kind of… I'm fine to use it because I just take stuff off it, so with WhatsApp, my… understanding of it is that it is secure end-to-end etc., so the stuff I give out there, if that trust is broken, that would just…. I would just stop using it… If I knew people could read it, or if they could take it, I would just be like: "That is completely broken!".'*

These comments from Liam and Ken encapsulate the links many participants drew between their trust towards SMS and digital services, and the security of these services when handling their information. Throughout focus groups, distrust towards SMS providers was particularly apparent. This appeared in line with data from polls suggesting trust towards SMS providers is significantly lower than manufacturers of others technologies, such as credit cards and smartphones (Rainie, 2018). Moreover, in a similar poll of SMS users' attitudes towards providers, Facebook had the worst reputation, with 58% of users reporting little-to-no confidence in the company protecting their information and data (Gallagher, 2018). The low trust reported in these findings likely reflected the fallout from the Cambridge Analytica scandal, where massive amounts of user data was inappropriately shared with third party organizations (Cadwalladr & Graham-Harrison, 2018). Likewise, distrust of SMS within our focus groups was also possibly informed by the recency of the Cambridge Analytica scandal, as the story had featured frequently in the news in the 3-4 months prior to the focus groups we conducted. As such, it's difficult to rule out the extent to which the Cambridge Analytica scandal has colored participants' views, and it's possible that this event enhanced their distrust and skepticism of SMS and digital services. Nonetheless, previous research has illustrated that system security is also a significant factor in users' trust towards other digital services. In a sample of Bosnian university students, acceptance of online banking services was informed by the perceived security of the system, as well as the users' perceived ability to use the system (Özlen and Djedovic 2017). Therefore, trust towards SMS and digital services appeared to be closely linked to the security of the platforms providing the service. However, in the case of SMS, there was also consideration for how information can be spread by other users on the platform.

2.4.1.2 Other Users

The role of *Other Users* appeared as a subtheme within the theme of Information Security, in which participants' trust towards SMS extended to include other users on these sites. While most participants expressed an awareness over the amount of control they had over the information they shared with these platforms, they also acknowledged that their personal information was vulnerable when shared by other users.

Patrick: *'I think you have got so much [control] over the privacy settings that you can set: No one can find me, or only my friends can view this, […] so you've already chosen to trust those people, so I think it's more like the human trust than trusting an app.'*

Patrick's comments highlight participants' view of SMS as platforms in which information security is enforced by both providers and users. While providers often give users access to privacy and security settings, when sharing information publicly, participants' trust towards SMS extended to include their evaluations of the other users within their social networks. For some participants, this meant restricting what they shared so that it could only be seen by other users within their close friendship circle.

Claire: *'Well I always keep it on private mode or whatever, like I wouldn't just have it out there, but I don't really think about it on a regular basis, in terms of what I'm sharing with my friends.'*

Claire's comment suggests that for some users, if they can trust the other users in their SMS networks, and limit the access of users they do not trust, then they may be more likely to share information on the platform. For other participants, these restrictions were based more on the content itself.

Cher: *'I just don't put anything out there that I wouldn't want […] the internet to know.'*

Comments from Claire and Cher highlight beliefs that SMS providers have limited control over publicly shared information, and emphasizes the users' need to self-regulate the information they share. Thus, for some participants, trust in SMS as a platform was informed by a shared responsibility for their information between both the user and the provider. This makes sense given users will be more likely to share information on SMS when they perceive themselves to have more control over their information and privacy (Jung, 2017). For example, Benson and colleagues (2015) report that SMS users' feelings of control over their personal information is linked to how much personal information they disclose. Thus, participants' trust towards digital services was informed by the perceived security of the platform, and in the case of SMS these evaluations extended to include the role of *Other Users* within their friendship networks. However, for some participants their engagement with SMS and digital services were not informed by feelings of control.

2.4.1.3 Lack of Control

While some participants believed that they had control over their data and information, other participants felt that they had limited control over how their information could be shared. The subtheme *Lack of Control* illustrates how some participants' felt coerced into using these services. For some participants this stemmed from the ubiquity of digital services in modern life.

Melanie: *'You're almost forced into trusting […] everything I do, if I like buy something online, if I transfer money to someone, like everything is in a system somewhere.'*

Melanie's suggestion illustrates the rising pervasiveness of digital services, in which users may feel engagement is a growing requirement for many aspects of life in the modern world. Her claim of being 'forced into trusting' characterizes her engagement with these services as being motivated more by compliance than by trust. This was echoed by other participants, who found it difficult to deal with the Terms and Conditions associated with these services.

Adam: *'Gmail […] there was a period of time when I was trying to access the emails, like I had to send like an email quickly, and it had like a notification that they had applied all the privacy stuff, […] and it was like it was like a huge document, and I had no other chance, other than clicking apply, because […] even if I don't agree it's my Gmail…'.*

Thus, within the Information Security theme, individuals' feelings of control over the information they shared appeared to influence their trust towards the service. However, some participants expressed a lack of control in their engagement with digital services, which may potentially limit their trust in these services. Similar views were also expressed by other participants, in which they felt engagement with SMS was a requirement for 'fitting in' within social groups.

Lana: *'I think its because [for] so many people, and its so useful… that now its expected, that you should be using it, not as a something of leisure, but as a work thing.'*

Carlos: '*It's one of these things where it's reached that kinda critical mass, where everyone has them, [...] it does make things a lot easier on you.'*

Jane: *'I guess we felt it was necessary because people kept like asking "why can't we contact you online?".'*

These views are somewhat unsurprising, given SMS continue to be one of the main reasons we use other technologies, particularly smartphones and tablets (Poushter, 2016). The ubiquity of SMS may be a powerful motivator for prospective users. However, on one hand this may not necessarily be a bad thing itself, as Park and colleagues (2013) report, increased usage of smartphones and SMS are positively associated with increased social capital; the individual's ability to maintain social interactions with others. At the same

time, while SMS provide users with a platform to maintain more friendships, this may not automatically translate into long-lasting, high quality friendships. Pittman and Reich (2016) report greater uptake in SMS use has coincided with greater feelings of loneliness in many Western countries. While Fear of Missing Out (FoMO) is a term that is often used playfully in advertising and media, research suggests FoMO may be a legitimate driver for SMS engagement amongst users. Roberts and David (2020) argue that while FoMO is typically associated with negative consequences for the intensity of SMS use, under the right circumstances FoMO can positively influence users' wellbeing, if it motivates them to use SMS in a manner that strengthens social connections. Ultimately, while *Lack of Control* may damage user's trust towards the platform, this may not necessarily be a bad thing if this control is traded for convenience and the increased social capital that digital services and SMS can offer to users.

The theme of Information Security broadly covered how participants' trust towards SMS and digital services was informed by how secure they believed their information and data were, when using these platforms. While the security of the platform itself seemed to be a primary determinant in how much users engaged with the service, they also considered the effect that other users can have, at least in the case of SMS. Moreover, participants also described instances where they felt their use of SMS and digital services were less informed by choice, and instead driven more by necessity.

2.4.2 Theme 2: Service Transparency

Throughout focus group discussions, most participants demonstrated some awareness of the information collection processes used by SMS and digital service providers. However, there was a greater deal of variance in participants' understanding and interpretation of these practices. This appears to result from a lack of transparency in the information collection methods used by SMS and other digital services. In the theme of Service Transparency, participants' trust appeared to be characterized in three different subthemes. Firstly, participants showed differing levels of *Understanding* regarding the data collection practices used by SMS and digital services. Some participants demonstrated a more nuanced understanding, whilst others did not fully understand the ways in which their data was collected. Similar to this, the subtheme of *Apathy* captured how some participants placed little-to-no value in how their data was collected, wherein their trust may have limited influence on their use of the service. Lastly, the subtheme *Unclear Intentions* captured how some participants' distrust of SMS and digital services stemmed from the unclear intentions of the service provider when collecting their data.

2.4.2.1 Understanding

Across the focus groups, participants showed various levels of *Understanding* regarding how their data was collected when using SMS and other digital services. For example, some participants like Anna, had a strong grasp of these procedures, and considered how different information may be collected by the different types of digital service they used.

Anna: *'So there's a distinction right, with LinkedIn and Facebook, they're targeting you, they're selling your data to advertisers so that they can specifically target you, whereas Netflix is using it to build their AI, so is there a distinction there? Because, with Facebook and LinkedIn you can change your privacy settings, so you don't share as much data with them, and with Netflix you can't, because no one is getting that data except for Netflix.'*

Anna's comments illustrate how some participants contextualized their trust towards digital services by considering the type of digital service they were using, along with their perceived intentions of the company providing the service. Similar awareness was also demonstrated by other participants.

Patrick: *'the whole GDPR, its… its not actually finding you, the person, its finding you, 25–30-year-old, male who likes football, sport etc. its not actually seeing you as a person, its about seeing you as… the other data […] they're not targeting you specifically'*.

Anna and Patrick's characterizations highlight the ways in which some participants went beyond thinking about themselves as users sharing information within a platform, and instead considered the role of the companies that provide these platforms. Tufekci (2015) suggests that SMS users are more likely to share personal information when they know how it is being used by the service. Similar findings were also reported by Benson and colleagues (2015), who report that users were also more likely to disclose personal information when they have greater awareness of how this information is used by SMS providers. Thus, for individuals like Anna and Patrick, their trust towards SMS and digital service providers, and the information they subsequently share with these services, may be framed in their understanding of how their data could be used by these companies. However, not all participants demonstrated the same level of understanding for how their data was being used by SMS and digital service providers.

Barbara: *'Yeah its not like a trade-off because you don't really understand what you are giving up. I guess, Like I don't know what… it never crosses my… maybe when people… you know when it comes up with a cookie thing, those, it will come up a statement, that's the only time I'll think for a second "Oh someone might have some information of mine",*

but I don't know what they have, what they use it for, what the benefit is of having information about me.'

Barbara's comments illustrate how the information collection processes of SMS and digital services are not always easily understood. However, this lack of understanding was even shared by some of the participants who considered themselves as knowledgeable about these practices.

Ken: *'I'm reasonably informed about stuff, I… I would say every year in the last 10 years, I have been amazed continually at stuff I didn't know, about whether its about cookies to start with, whether its about Geotargeting in your phone, whether its about a fact that, maybe anytime I log into WIFI now, all of my data is basically accessible, like… my boardroom at work, they're not allowed phones in there'.*

When taken together with Anna and Patrick's comments, these comments highlight a subtheme of *Understanding* within the Service Transparency theme. Users' trust towards SMS and digital services may be inherently limited by a user's ability to understand how their information is used by the service provider. This in line with Tufekci (2015) who also suggests that a lack of transparency can damage users' attitudes towards a service, and make some users less likely to share personal information. Thus, while some users' trust towards SMS and other digital services may be partly characterized by their beliefs about the provider's intentions for their data, for other individuals there was limited understanding of what companies were actually doing with their information.

2.4.2.2 Apathy

Within the Service Transparency theme, there was also a subtheme of *Apathy*, which was marked by some participants' general disinterest in how their information was used. While most participants expressed an awareness of their information being collected by SMS and digital services, for some there was a lack of interest in how this information may be used by the service provider. This *Apathy* appeared to stem from both a lack of transparency in the service providers' information collection procedures, and personal disinterest in the value of their information.

Bean: *'Yeah information mining […] I assume that that stuff is going on, yet I use the platforms anyway, because, I dunno it's like good enough for me, I guess I assume I'm like a bit insignificant and like a bit boring… I'm just average.'*

Bean's comments characterize a more passive form of participants' trust towards digital services, in which she is aware of the data collection motivations of these providers, yet continues to use their platforms and services. Similar comments from Barbara also highlight how some participants placed limited value in their information.

Barbara: *I don't think anyone's interested in my information […] so I don't think about that'.*

From this perspective, some participants' trust towards SMS and digital services may be influenced by a lack of value attributed to their personal information and data. Such views may help to explain why many users continue to use SMS and digital services after high-profile data breach scandals. In the wake of the Cambridge Analytica scandal (Cadwalladr & Graham-Harrison, 2018), Facebook's reported daily user activity and overall stock valuation decreased significantly (Solon, 2018), suggesting that users' trust was damaged by these events. However, this event did not lead to the collapse of Facebook as a platform, as users continue to engage with their service. This may imply engagement does not appear to be entirely informed by the perceived trustworthiness of the companies that provide the platforms. Parallels with this can also be seen in the continuous growth in Twitter's userbase (Kastrenakes, 2021), despite the platform facing continued pressure to halt the spread of misinformation in their service (Oyeyemi et al., 2014; Jin et al., 2014; Hiar, 2021). Thus, within the theme of Service Transparency, users' understanding of how their information and data are used by service providers could make them more likely to share personal information with the platform. Somewhat paradoxically, the *Apathy* subtheme captured how some participants' a lack of understanding, or disinterest in the service provider's information handling policies, could also lead them to place limited value in their personal information, and subsequently more likely to share their personal information.

2.4.2.3 Unclear Intentions

There was also evidence of distrust towards SMS and digital services, which appeared to stem from a lack of transparency from these service providers. When contemplating trust, some users considered the intentions of the companies that provide these digital services, and had difficulty trusting them if they believed these companies had malicious intentions. For example, while Barbara admitted to having limited understanding or interest in how her information and data were used by SMS and digital service providers, she also appeared to distrust the platforms as a result of this limited transparency.

Barbara: *'I speak about something and the next thing you know it's on my Facebook [advertised] […] they're listening'.*

Barbara's comments illustrate how this lack of transparency could manifest as distrust for some participants, with particular concern for the targeted advertising methods that these services employ (Lipsman et al, 2012; Alhabash et al., 2017; Lee et al., 2018). While SMS such as Facebook and Instagram utilise targeted adverting based on

users activity within their apps, there is currently no evidence that they access users' microphones to monitor their conversations outwith these apps (Tulek & Arnell, 2019). For other participants, their view of these information gathering practices was more cynical.

Adam: *'…there's no actual aim to do value or to do anything that's benefiting the users, it's mainly just getting them with, with any sort of kind of bubble, to keep engaged…'*

Comments from Adam and Barbara illustrate how some participants believed that digital service providers sought to exploit their users as a way to generate revenue. This combination of low transparency and the use of targeted advertisements appears to be potentially damaging to trust and corresponds with previous findings in the literature. In a study examining the influence of targeted advertisements in social media (Jung, 2017), users concern about privacy increased when exposed to advertisements that were highly personalized towards them, making them feel their information was being actively tracked. Whilst advertising revenue is undoubtedly a core financial source for these companies (Alhabash et al., 2017; Lee et al., 2018), the ways in which users' data is collected and used are open to interpretation, and some users may interpret a lack of transparency as insidious intentions. However, while most participants expressed an awareness of these targeted advertising practises, not all participants perceived them as malicious.

Melanie: *'Like how they monitor what websites you go on? I feel like it doesn't bother me as much as it should, like I think that a lot of people get freaked out, like this whole media thing of like phones listening to you and all that, but I just don't really have anything to hide […]'*

Thus, awareness of information gathering and targeted advertisements does not appear automatically foster distrust amongst all users. As Melanie's comments illustrate, despite not fully understanding how her data is used by these companies, for some users these practices may be perceived as being benign. Kim and colleagues (2019) report that higher trust towards a digital platform can make targeted adverts more acceptable to viewers. Taken together, this would suggest that targeted advertisements may not intrinsically damage trust towards SMS and digital services themselves. Rather it appears that their influence on trust also depends on the value that users' attach to their personal information and data, and their general understanding of how this information could be used. Ultimately, the theme of Service Transparency illustrates a complex relationship between the transparency of SMS and other digital services, their use of targeted advertising, and the user's understanding and interest in how their information is used.

2.4.3 General Discussion

This study explored how people characterise their trust towards technologies, such as online digital services and social media platforms. Participants trust towards these services were characterised within two main themes: *Information Security* and *Service Transparency*. *Information Security* related to participants' thoughts and concerns about sharing their information with these technologies, and how securely their information would be stored by the companies providing these services. This echoes previous research, in which users' trust towards online services was informed by the perceived security of these platforms, as well as the amount of control users felt that had over their information (Özlen and Djedovic 2017; Benson et al., 2015; Jung, 2017). Thus, companies providing digital services may benefit from greater trust from users by providing clearer insights into how they securely store users' information, as well as providing users with increased control over the information they share with these services. *Service Transparency* captured participants' beliefs about what these companies did with their information. Previous research has also suggested that if users have a better understanding of how their information is used, they will be more likely to share it with these companies (Benson et al., 2015; Tufekci, 2015). If a digital service relies upon users willingly sharing their information with the platform, these companies way benefit from being more transparent with how this information is being used, and how this may affect users, through secondary technologies such as targeted advertising. Users may also be better able to calibrate their trust towards these services if they can become more aware of how technologies like targeted advertising work.

While this research has focussed on trust towards digital services and social media platforms, some of our findings may be generalisable for understanding users' trust towards other types of technology. Many types of autonomous systems, such as those in used in healthcare settings, may require access to users' personal information (Challen et al., 2019; Sujan et al., 2019; Goddard et al., 2014; Lyell et al., 2018). Trust may be gained more readily from users in these situations if the system is transparent with how this information is used. Similarly, we have characterised how users may seek to have control the information we disclose to these services, and control/authority is also highlighted within research involving other technologies (Abraham et al., 2016; Muir 1987; Goddard et al., 2014; Lyell et al., 2018; Chavaillaz et al., 2016; Chavaillaz et al, 2020). Both themes also show participants considered the reputations of the companies providing these services as factors relevant to their trust. Similar research has also suggested that trust

towards automation may be informed by the users' trust and attitudes towards the company that provided the automation, as well as the human designers that made them (Culley & Madhavan, 2013; Morgan-Thomas & Veloutsou, 2013; Celmer et al, 2018). Thus, our research also supports existing evidence suggesting that users' trust towards technology is partly informed by the reputations of the companies providing the technology.

2.5 Conclusion

Our study sought to understand how individuals characterize their trust towards social media and other digital services. Views from participants illustrated the many ways in which participants considered the trustworthiness of these services, captured by two main themes of *Information Security* and *Service Transparency*. In the theme of *Information Security*, participants' trust appeared to be characterized by the security of the platform they were using, with particular concern for how easily their information could be accessed by other parties. Participants' trust was also based upon how much control they believed they had (if any) over the information they shared with the service. In the case of SMS, participants also factored in the role other users may play in sharing their information, wherein trust towards the platform is also based upon their trust in the other users they may interact when using these services. In the theme of *Service Transparency*, participants trust also appeared to be characterized by their understanding of how their data is collected by the service provider. Some participants benefited from greater knowledge of how data is collected and used within digital economies, which informed their trust in digital services. Conversely, other participants demonstrated limited understanding of these practices, and based their trust on speculation and/or disinterest in the perceived intentions of these digital service providers. Participants' trust was also further informed by the value that they attached to their data once it was harvested, with some participants attributing little value to the information they share with these services. Collectively, participants' trust was characterized by a variety of expectations and assumptions regarding the digital services themselves, as well as the company that provides these services.

2.6 Focus Group Questions

2.6.1 Topic: Social Media

1. If you use it, do you feel using social media is enjoyable?
2. Do you feel that social media improves society, and if so why?
3. Is having an online presence/identity a necessary requirement for the modern world?
4. How important is trust for using social media?
5. Is it necessary to trust a social media platform for you to use and rely upon it?
6. Are the benefits of social media (such as maintaining social networks) an acceptable trade-off for allowing companies access to your information?

2.6.2 Topic: Information security

1. Could I ask you to write down some companies that come to mind when talking about data security – and specify whether you think they have a good or bad reputation (Written)
2. Do these reputations influence how likely you are to use their services?
3. What does it mean to you, to feel safe online?
4. Do you feel that your data and information is safer with a human or a computer, and why?
5. If you had to pick one, would you prefer that Public Governments or Private Companies had more access to, and control over your information?

2.6.3 Topic: Technology

1. If I were to mention the term Artificial Intelligence, what would you say is the first thing that comes to mind? (Written)
2. Can you give some examples of where you think artificial intelligence is currently being used, if any?
3. If a computer was described as having a form of Artificial Intelligence would that change your perception and expectations of it?
4. When using a machine/computer, if it makes an error/mistake, how does that affect your ability to trust it?
5. Do you think we can trust machines the same way we trust people?
6. Do people think that Artificial Intelligence and other digital technologies can be a good thing for society?

Chapter 3: Calibrating Trust Towards an Autonomous Image Classifier

Publication: https://doi.org/10.1037/tmb0000032

3.1 Abstract

Successful adoption of autonomous systems requires appropriate trust from human users, with trust calibrated to reflect true system performance. Autonomous image classifiers are one such example and can be used in a variety of settings to independently identify the contents of image data. We investigated users' trust when collaborating with an autonomous image classifier system that we created using the AlexNet model (Krizhevsky et al., 2012). Participants collaborated with the classifier during an image classification task in which the classifier provided labels that either correctly or incorrectly described the contents of images. This task was complicated by the quality of the images processed by the human-classifier team: 50% of the trials featured images that were cropped and blurred, thereby partially obscuring their contents. Across 160 single-image trials, we examined trust towards the classifier, while we also looked at how participants complied with the classifier by accepting or rejecting the labels it provided. Furthermore, we investigated whether trust towards the classifier could be improved by increasing the transparency of the classifier's interface, by displaying system confidence information in three different ways, which were compared to a control interface without confidence information. Results showed that trust towards the classifier was primarily based on system performance, yet this also was influenced by the quality of the images and individual differences amongst participants. While participants typically preferred classifier interfaces that presented confidence information, it did not appear to improve participants' trust towards the classifier.

3.2 Introduction

The success of new technologies is dependent on whether they are accepted by the end user. Our understanding of how users accept new technologies has developed over time, the initial Technology Acceptance Model (TAM) put forward by Davis and colleagues (1989) was heavily centred on the perceived usefulness and perceived ease of use of the system, as the primary determinants for technology acceptance. More recently, extensive work by Venkatesh and colleagues (2012) has sought to develop upon earlier iterations of TAM by integrating further, more diverse determinants of acceptance, such as system price, the user's habits, and even the hedonistic pleasure gained from using the

system (Venkatesh, 2015). This suggests that innovation alone is not enough for new technologies to be successful, and that there is a myriad of psychological, social, and environmental factors that inform the ultimate acceptance of technology.

While the successful adoption of new technologies is tied to users' acceptance of them, the users also need to learn to use the technology correctly. Just because someone accepts a new technology, it does not automatically follow that they will use it appropriately. This is particularly the case with autonomous systems, which are technologies that use Artificial Intelligence (AI) to undertake tasks with a degree of independence from their user. As these autonomous systems become more advanced, their capacity for complex tasks also increases, yet with this the opportunity for errors increases too (Parasuraman et al., 2000). As such the success of autonomous systems also relies upon appropriate trust from their human user, to ensure these systems are used correctly. Ideally operators' trust will be calibrated to reflect the actual performance capabilities of the autonomous system, ensuring they do not distrust a functional system (too little trust), or mistrust a dysfunctional system (too much trust) (Muir, 1987; Parasuraman & Riley, 1997). In this study, we sought to understand how humans calibrate their trust towards an autonomous image classifier system (AICS).

3.2.1 Autonomous Image Classifier Systems

AICS are technologies that can independently classify the contents of image-based data, using advances in deep learning and convolutional neural network research (Chan et al., 2015; Howard, 2013). A major advantage of AICS is that they can process large quantities of data quickly and independently, thereby reducing demand on human users. For example, in the UK, London's Metropolitan police force are interested in using AICS to help process digital forensic evidence, to reduce their officers' workload and limit their exposure to graphic content (Murphy, 2017). Moreover, AICS can be trained to distinguish specific, highly complicated patterns and features: an AICS was recently able to identify breast cancer with an accuracy comparable to human experts (McKinney et al., 2020). AICS can also be used in lower stakes settings, for example the popular app 'PlantNet' can provide users with classifications for images of plants and flowers that they encounter (Goëau et al., 2014). Even though these applications are impressive, the performance of AICS can reflect the expertise and potential biases of the engineers who design the systems, as well as the quality of the dataset used to train their algorithms (Danks & London, 2017; Rudin, 2019). Thus, AICS are vulnerable to errors and will require appropriate trust from human operators. This is particularly important, given the potential application of AICS in a wide variety of settings, where AICS may be responsible for

supporting high stakes decisions. Thus, we sought to examine how users calibrated their trust towards an AICS, and how this trust translated into compliance with the system's decisions, when completing an image classification task. By doing so, we provide an insight into trust specifically towards AICS, which we hope will benefit the design and deployment of AICS in real-world settings, while also providing further insights for the wider trust-in-automation literature.

3.2.2 Understanding Trust Towards Automation

Across the literature, trust-in-automation has been studied in a wide variety of human-machine teams, and arguably has most commonly been studied with autonomous vehicles (Jing et al., 2020). When considering how technology is used in different human-machine teams, Larson and DeChurch (2020) make a distinction between technology and agents. Technology is something that is used by teams to achieve their goals, much like a tool, while agents fill a distinct role within the team which goes beyond mere augmentation, and inherently improves the team's performance as a result (Larson & DeChurch, 2020). For agents, they also draw a distinction between robots, which are agents with embodied physical characteristics, and AI which are disembodied agents that perform tasks that traditionally require human intelligence, such as visual identification and decision-making. Trust has previously been studied with both robot-based agents (Selkowitz et al., 2017; Desai et al., 2013), and with AI-based agents, such as automated software repair systems (Ryan et al., 2019), virtual cognitive agents (Hertz & Wiese, 2019; de Visser et al., 2016), and decision support systems (Sauer et al., 2016; Yu et al., 2019; Zhang et al., 2020). Regarding AICS, these systems most closely align with the examples of AI-based agents. It should however be noted that within our experimental design, we afforded the AICS limited agency, as human users supervised each classification decision, with the authority to overrule each one. Whereas in real-world applications, AICS may be employed as agents with greater autonomy when working within teams.

We interpreted trust towards an AICS through the lens of Hoff and Bashir's (2015) model of trust towards automation, which separates trust into three broad layers. Dispositional Trust relates to stable human-centric factors, such as culture, age and personality traits, which inform users' general disposition towards technology. This would reflect the users' attitudes towards the AICS, and more broadly technology in general. Situational Trust relates to fluctuating human-centric factors, such as mood and attention, as well as environmental factors, such as task difficulty, workload and organizational setting, which can all vary over time. We believe that when using an AICS, a significant factor for operators' trust would be the quality of images being processed, which could

increase the difficulty of system classifications, particularly if the operator feels they could easily classify the images themselves. Finally, Learned Trust is split into two separate sub-layers: Initial Learned Trust that reflects the user's historical experience of similar systems, and the reputation of the current system, while Dynamic Learned Trust reflects their ongoing experiences of working with the system. When working with an AICS, Learned Trust will likely be informed by the users' ability to interpret the system's decision-making, particularly if the image is difficult to classify. Additionally, in industrial applications, operators may have previous experiences with other AICS, which may inform their trust towards newly introduced systems. Hoff and Bashir (2015) suggest these three layers of trust combine to ultimately inform how users rely upon the autonomous systems during collaboration, which would be crucial for appropriate use of AICS. Therefore, when investigating trust towards an AICS, we created experimental manipulations that were consistent with Hoff and Bashir's (2015) model and contextualised our hypotheses and subsequent findings within their theoretical framework.

3.2.3 System Performance

Hoff and Bashir (2015) demonstrate the complex relationship between human, mechanical, and environmental factors that combine to inform trust towards autonomous systems. However, their model stipulates that when interacting with automation, system performance is the central modulator of trust towards automation. In this vein, Yu and colleagues (2019) reported close relationships between perceived system accuracy, trust, and reliance upon an automated fault detection system, and demonstrated that users will modulate their trust and reliance in response to system performance. Thus, when collaborating with an AICS, we anticipated system performance, defined as the classifier's ability to correctly label the contents of images, will have the biggest influence on trust: *(H1a)* System performance, whether the classifier's label correctly describes images, will have the strongest influence on trust towards the classifier.

3.2.4 Image Clarity

While system performance should be the main driver of trust towards the AICS, the classifier's performance itself is likely to be dependent upon the quality of images being processed. Hoff and Bashir's (2015) Situational Trust encompasses factors which make tasks more difficult to accomplish, and we believe image quality would be a particularly important factor within the context of AICS use. When images have lower clarity, through factors such as occlusion and blurring, the contents of the image may be harder for human users to identify. Moreover, when an AICS processes lower clarity images, the system's

performance is also likely to be harder to evaluate, given the increased uncertainty of the contents of the images, which may itself impact upon trust towards the classifier. Thus, when working with an AICS the quality of the images processed could be considered as an environmental factor, given the operator may have limited control over image clarity. A similar issue was explored in a study by Merritt and colleagues (2013) involving trust towards an automated baggage scanner where trust towards the scanner was affected by the difficulty of the task. Specifically, trust was lowest in blocks where the scanner's performance was considered as 'obviously poor', and highest when 'obviously good', given the presence of weapons was made relatively obvious to participants. However, in the more difficult, ambiguous block, where the contents of luggage were cluttered, trust was found to be lower than the 'obviously good' block, yet higher than the 'obviously bad' block, illustrating the effect of task difficulty. Similar findings were reported in another study that involved an automated letter detection aid, in which participants were more likely to accept the system's advice in trials with higher difficulty (Schwark et al., 2010). This suggests that the difficulty of the task facing human-machine teams may influence how human users interpret and use automated system advice. Of course, the influence of task difficulty is likely to vary between autonomous systems, as different systems will be employed in different occupational settings, with varying consequences associated with system errors. Nonetheless, we anticipated that the relationship between system performance and trust towards the AICS would be modulated by the quality of the image being processed: *(H1b)* Image Clarity will significantly interact with system performance when predicting trust towards the classifier. With unclear trials, trust will be lower when the classifier is correct, and higher when the classifier is incorrect, illustrating participants' uncertainty about the classifier's performance.

3.2.5 Individual Differences

Trust towards an AICS could also be influenced by the operator's cognitive understanding of the system and task, which can be prone to biases intrinsic to each individual (Israelsen & Ahmed, 2019). Some examples of these biases include: Automation Bias, where automation performance is perceived as inherently superior to human performance (Goddard et al., 2011); and Perfect Automation Schema, where individuals may believe that automation is almost always perfectly reliable (Dzindolet et al., 2002). These biases reflect differences in trust stemming from the experiences of individual human users. Hoff and Bashir (2015) characterise biases towards trusting machines as a form of Dispositional Trust, which are relatively stable over time, and reflect users' tendencies independently of context. In order to understand how human-centric

factors influenced trust towards the AICS, we considered each participant's score in the Propensity to Trust Machines Questionnaire (PTMQ) (Merrit, 2011), as a form of Dispositional Trust. PTMQ scores can be used to characterise each user's predisposition towards trusting technology, in which higher scores represent higher self-reported tendencies to trust new technologies. The use of PTMQ was highlighted in the study by Merritt and colleagues (2013), which showed individuals with higher PTMQ scores had higher trust towards the automated baggage scanner when it processed luggage with cluttered contents, during the ambiguous performance block. This suggests that users with higher PTMQ scores were less likely to have their trust influenced by the difficulty of the task, even though the uncertainty of task success would make it harder to evaluate system performance more accurately. Thus, users' existing tendencies towards trusting machines may influence trust, even when environmental factors complicate their evaluations: *(H1c)* Participants with higher Propensity to Trust Machines scores will trust the classifier more when processing unclear images, where performance may be more difficult to evaluate.

3.2.6 Improving Trust Through Transparency

Trust towards autonomous systems may also be improved when system decision-making is made more transparent (Tomsett et al., 2020). For example, drivers reported greater trust towards a driving aid within an autonomous vehicle simulator when provided with explanatory feedback messages (Koo et al., 2015). The Situation awareness-based Agent Transparency (SAT) model proposes that autonomous system transparency can be improved by providing users with more detailed information that is relevant to system performance (Chen et al., 2014). Within the lens of the SAT model, human users may calibrate their trust more appropriately if the system provides more detailed information about its current task (Chen et al., 2014). Using the SAT model, Selkowitz and colleagues (2017) report increased trust towards an autonomous robotic squad member as it provided users with more detailed situational information, such as system motivations and predicted task outcomes. However, this trend was not apparent in the condition with the most information, implying there may be a limit to how much information is beneficial to users' trust (Selkowitz et al., 2017). Hoff and Bashir (2015) suggest that these design features which increase transparency can help users to understand the system's purpose and process when carrying out tasks, thereby improving the user's Learned Trust. Thus, we sought to understand if we could improve trust towards an AICS by making its decisions more transparent through displays of system confidence information (SCI).

3.2.6.1 System Confidence Information

SCI is a representation of system certainty when carrying out tasks and can benefit trust towards autonomous systems (Zhang et al., 2020). For example, SCI cues helped users to appropriately align their trust towards a navigational robot; lowering trust when confidence was low to accommodate poorer performance, and elevating trust when confidence was high (Desai et al., 2013). Similarly, Verame and colleagues (2016) report individuals were more likely to accept the decisions of an autonomous document reader when it displayed 'very high' confidence, compared to when displaying 'medium' or 'low' confidence. This suggests SCI may improve system transparency, and in turn users' trust and strategies for collaboration. However, there are a variety of ways that SCI can be represented within the interface of autonomous systems. Previous examples include confidence discretised into high/medium/low categories, represented with icons (Desai et al., 2013) or with text (Verame et al., 2016); as numerical probabilities (9/10 = high confidence) (Zhang et al., 2020); or visually through the color and opacity of icons (Selkowitz et al., 2017). Within the context of the SAT model, it is possible that more detailed forms of SCI would make AICS decision-making more transparent, and therefore be most likely to improve users' trust towards the system.

3.2.6.2 Complexity of System Confidence Information

Regarding systems specifically designed to classify image-based or text-based content, Ribeiro and colleagues (2016) suggest SCI could be displayed through a bar graph to illustrate the probabilities of the most likely options for each decision. Arguably Ribeiro and colleagues' (2016) suggestion presents SCI in a more transparent format than the previous examples above, as it provides the user with the system's confidence for the final decision relative to the confidence for other likely classification options. However, there is conflicting evidence surrounding the utility of bar graphs when conveying information, as evidence suggests they can be difficult to comprehend (Chaphalkar & Wu, 2020), and can lead to biases in readers' thinking (Godau et al., 2016). Contrarily, bar graphs have been considered useful when illustrating results with borderline differences, and reportedly require less time to interpret than raw data tables alone (Brewer et al., 2012). Therefore, we created three separate experimental interfaces that illustrated SCI in different formats and compared them against an interface without SCI (Control Interface). We adopted Riberio and colleagues' (2016) recommendation of using a bar graph to illustrate SCI (Graphical Interface), we also displayed SCI using text-based percentages (0-100%) (Numerical Interface), and lastly used color cues to represent SCI discretised into high/medium/low categories (Iconography Interface). Thus, we explored the benefits of displaying SCI within the AICS interface: *(H2a)* Relative to the control interface, the confidence

information presented within the experimental interfaces will improve overall trust towards the classifier.

We also sought to understand whether SCI would be more useful when the task difficulty increased, specifically when the classifier processes unclear images: *(H2b)* When processing unclear images, trust will be higher towards the experimental interfaces because they provide users with more information.

Lastly, we explored whether the addition of SCI in the experimental interfaces would increase users' workload, measured through subjective task load, and the amount of time participants spent in each trial: *(H2c)* When working with the experimental interfaces, participants' task load will be higher given interfaces with SCI present more information per trial.

3.3 Methods

3.3.1 Participants

74 participants (37F, 36M, 1 Non-Binary), primarily university students (Mean Age = 26.2, Min = 19, Max = 55), were recruited through the University of Glasgow's School of Psychology subject pool. All participants were compensated at a rate of £6 per hour for their time. 51% of participants considered themselves native English speakers. Ethical approval was obtained from the University of Glasgow, College of Science and Engineering ethics committee.

3.3.2 Design

We used a 2x2x4 within-subjects design where participants saw 2 levels of Classifier Performance (Correct, Incorrect) combined with 2 levels of Image Clarity (Clear, Unclear), within each of the 4 Interface-specific blocks (Control, Graphical, Iconography, Numerical). In each single-image trial (n=160) the classifier's label would either correctly or incorrectly match the image displayed, which was purposely made easy or difficult to evaluate due to the clarity of the image. The ordering of blocks was randomised, as was the ordering of trials within each block (n=40). The average participant took 17 seconds to complete each trial, and 12 minutes to complete each block.

3.3.3 Materials

3.3.3.1 Image Classifier

Participants interacted with an AICS based on the AlexNet image classifier model (Krizhevsky et al., 2012), which used MATLAB's Deep Learning and Image Processing Toolboxes (MATLAB ver. R2017a). AlexNet is a pretrained convolutional neural network, trained to classify objects within a 227x227-pixel net. To process each image, the file must first be resized to fit these dimensions, after which AlexNet is able to read the image. AlexNet can output a range of classifications and probabilities to illustrate its interpretation of images.

3.3.3.2 Classifier Performance

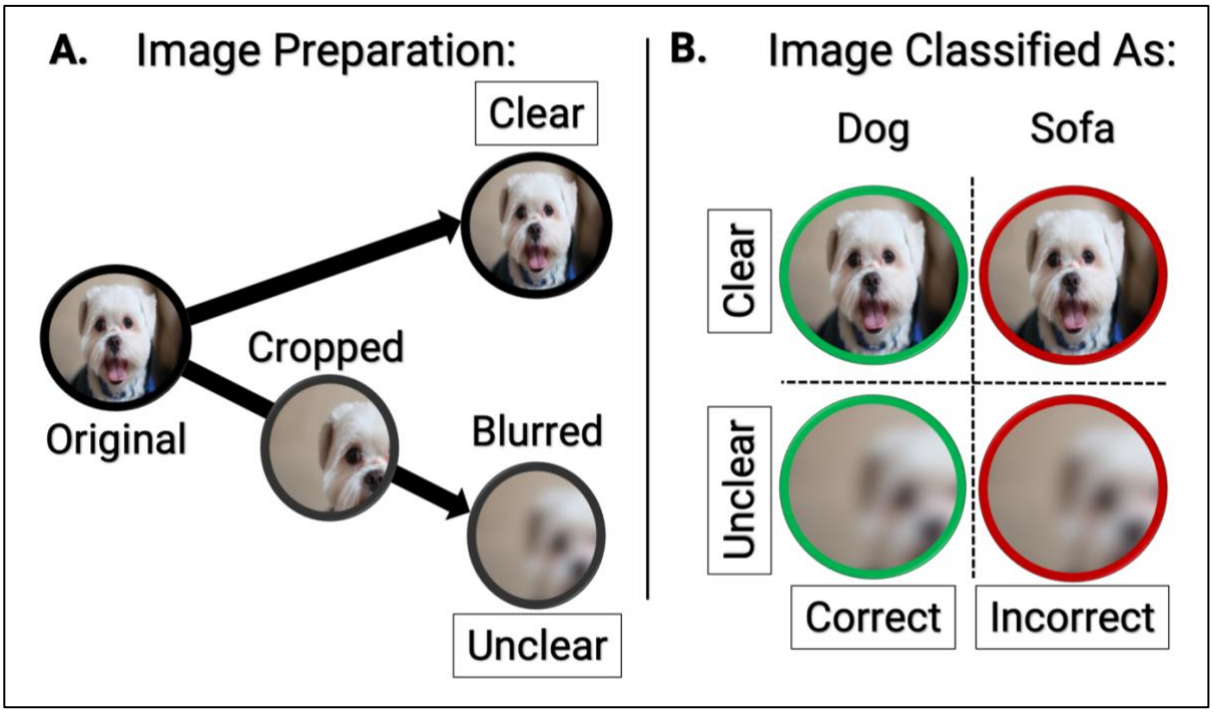
Participants viewed a series of 160 images selected from The Open Images Dataset V4 (OIDV4), (Kuznetsova et al., 2020). These images featured categories such as household objects, nature scenes, food items, vehicles, and animals. These were used to create four sets of 40 single-image trials, with each having 20 correct and 20 incorrect trials. The classifier's performance was considered as correct when AlexNet provided labels that appropriately matched the image's original label in OIDV4, otherwise

performance was considered incorrect. Classifier performance was intrinsically linked to each image; performance only varied between images.

3.3.3.3 Image Clarity

The contents within 50% of images was made unclear to make the classifier's performance harder to evaluate. These images were first cropped, to partially show their contents, and then overlaid with a Gaussian blur when displayed to participants (See Figure 1). Across all 160 trials, participants saw 40 trials of each combination of Classifier Performance and Image Clarity: Correct-Clear, Correct-Unclear, Incorrect-Clear, Incorrect-Unclear, which were evenly distributed and mixed across 4 sets of images. These sets were organized to ensure they contained the same quantity of categories (animals, vehicles, objects etc.), while the average classifier confidence was made similar in each set of images (Min = 49.5, Max = 53.6). Each set of images was randomly matched to an interface for each participant. Data associated with 1 image was corrupted during data collection, and therefore unusable (74 trials removed from initial 11840 observations).

Figure 1
Preparation of images



Note. Each trial featured a single image. Classifier performance was based on AlexNet's classification for the image (B), while Image Clarity was based on the quality of the image (A). Clear trials featured images with unobscured contents, while unclear trials featured cropped images that were overlaid with a Gaussian blur when presented to participants.

3.3.3.4 Image Classification Task

Participants used a mouse and keyboard to interact with the classifier's Graphical User Interface (GUI), built within the MATLAB app designer, (MATLAB ver. R2017a) (See Figure 2). The classifier's label for each image appeared in a box underneath the image, while participants could overwrite the classifier with their own label for each image. If participants did not understand the classifier's label, they could specify this with a small button beside the label. Additionally, if participants believed the classifier's label was wrong, yet were unable to provide a better correction themselves, they wrote "No" or "Don't Know" in their own user label box. Participants rated the classifier's performance on a visual analogue scale within the GUI, using 3 different interactive sliders corresponding with:

1.) **Image Familiarity:** Participants were asked to rate how familiar they were with the object in each image, insofar as they could identify what the contents of the image were. This was to assess how capable participants were at labelling the image themselves without the aid of the classifier. While most participants understood the instructions and meaning behind this slider, a few participants expressed confusion about the wording of

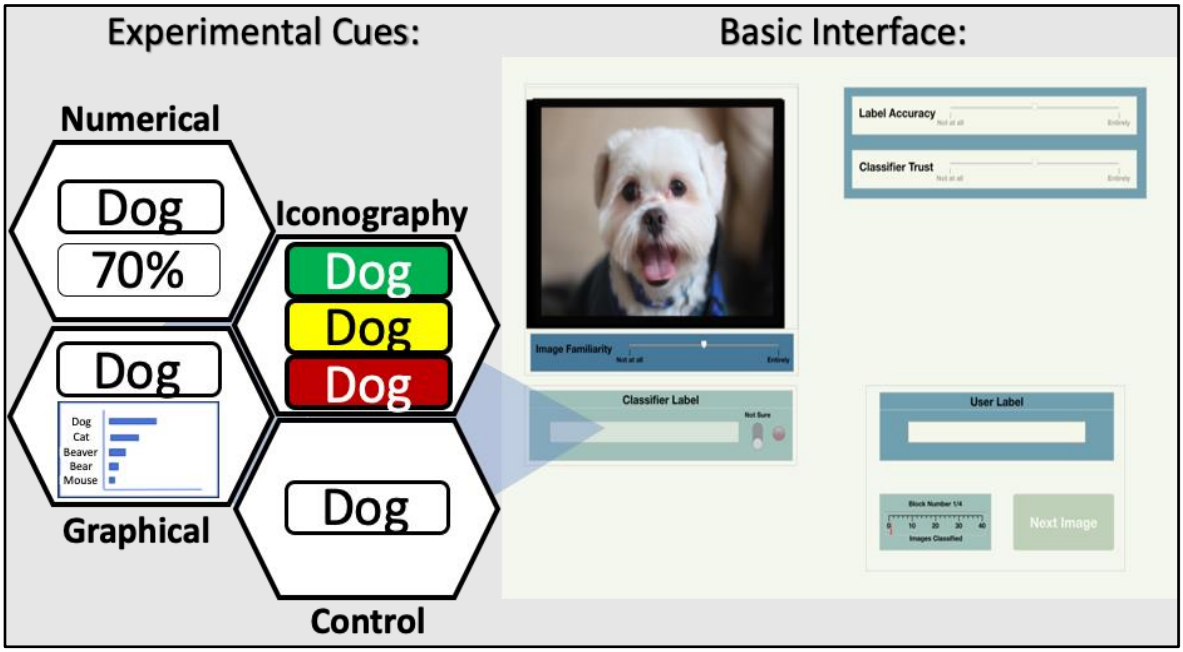
this slider, and this was subsequently renamed in Chapter 4 to ‘Labelling this image is easy’, which appeared to be a better explanation of this input for users.

2.) **Label Accuracy:** Participants rated how accurately they believed the classifier’s label described each image, and were asked not to consider the previous performances from the classifier in this evaluation. The intention of this input was to allow participants to rate the performance of the classifier on an individual level, by considering the performance only in the current trial.

3.) **Classifier Trust:** Participants then rated their trust towards the classifier. Participants were instructed that ratings of trust should represent their continuous interaction with the classifier throughout the experiment, and could be based on performances in previous trials. The intention of this input was to have participants appraise the cumulative performances of the classifier throughout the experiment.

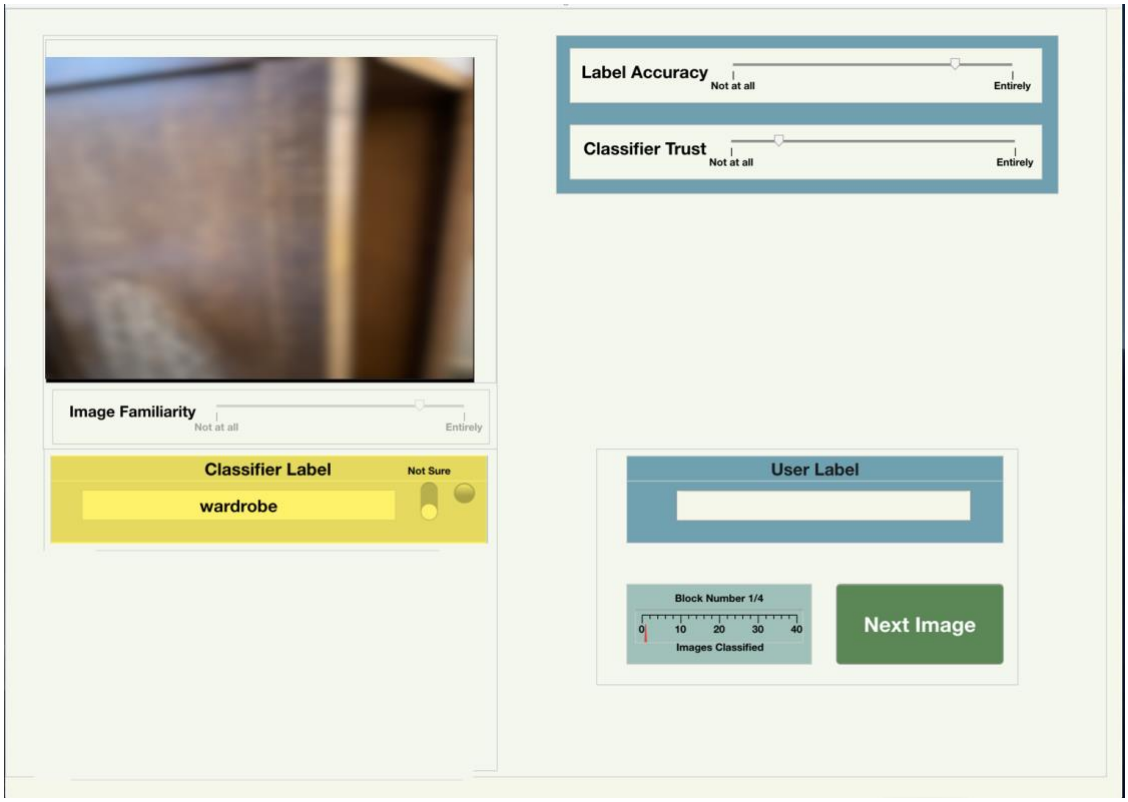
All sliders went from 0-100%, represented with visual anchor points of “Not at all” and “Entirely”. Data was collected from each slider after each trial and would reset to the midpoint (50%) between trials. Each slider would change colour (blue) to cue participants towards the rating they needed to provide next, guiding the participant throughout each trial. Compliance with the classifier was defined as trials where the participant did not overwrite the classifier’s label. Participants moved between trials by using the “Next Image” button, which only became active after all 3 sliders had been used.

Figure 2a
Interface Differences



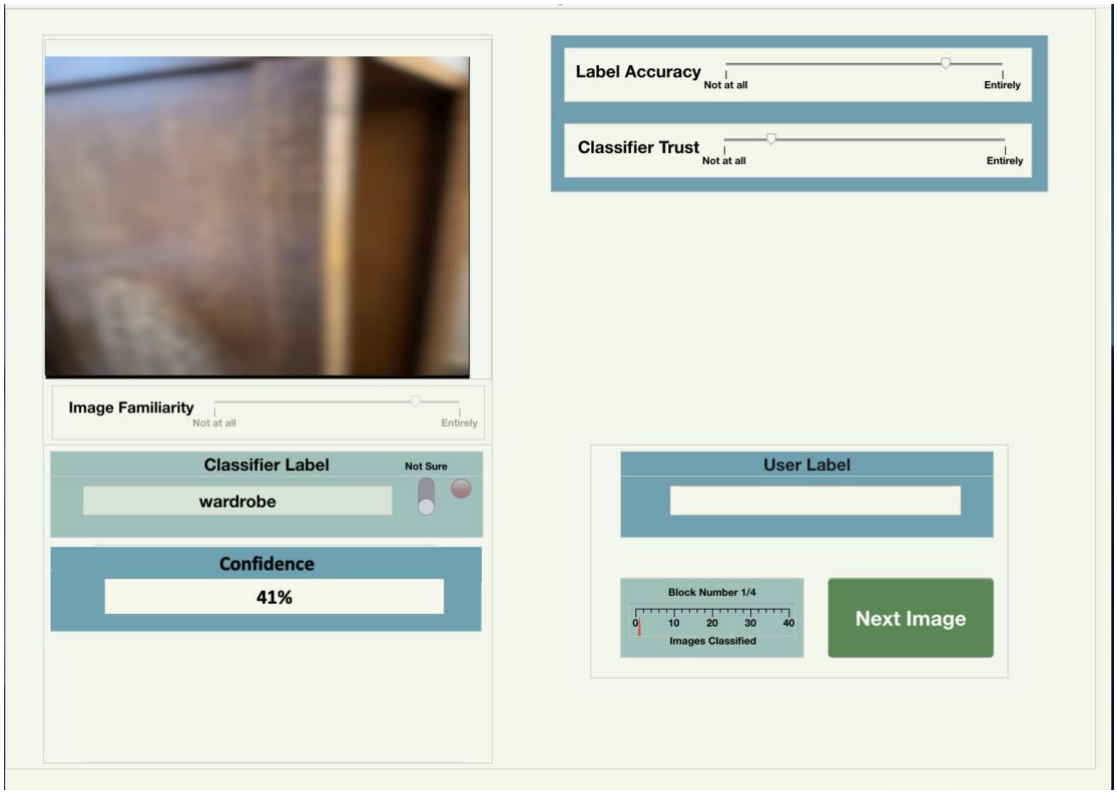
Note. All four classifier GUIs contained the same basic elements. Cues of SCI were only added to the lower left-hand side of the interface, to ensure visual similarity.

Figure 2b
Iconography Interface



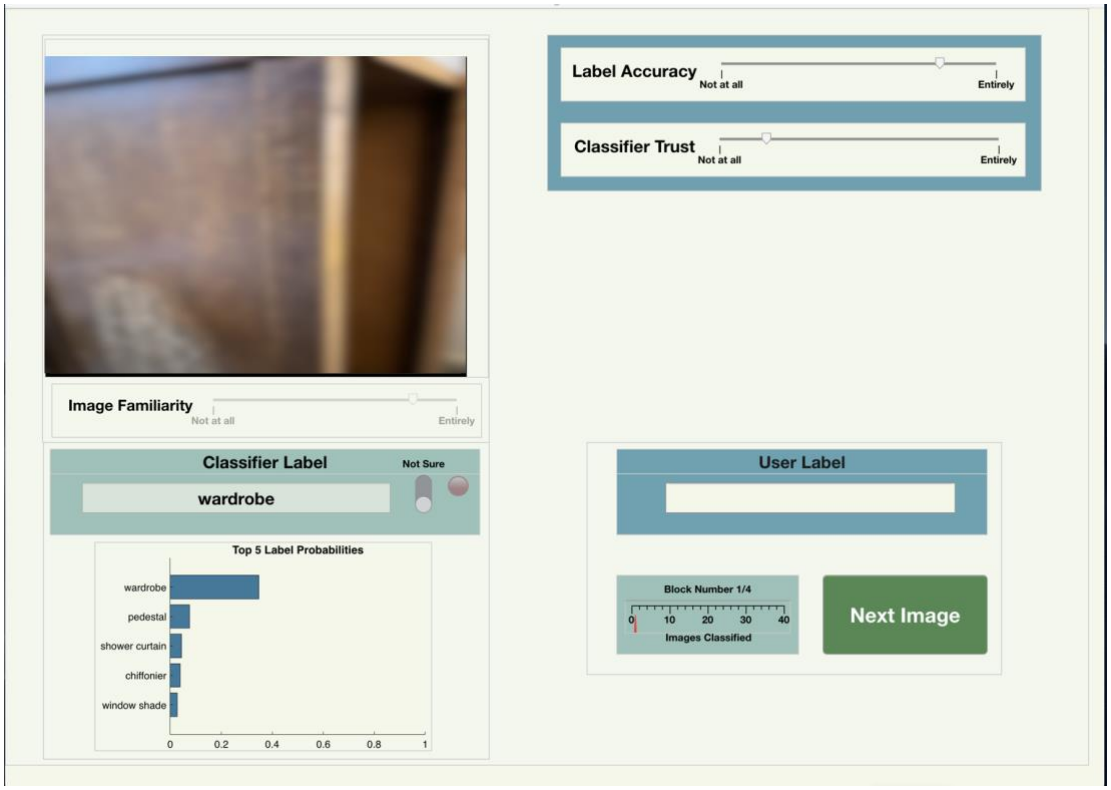
Note. The classifiers label in the Iconography interface changed colours to represent the classifier's confidence in the current label. Red represents low confidence, Yellow represents medium confidence, Green represents high confidence.

Figure 2c
Numerical Interface



Note. The classifiers SCI was represented as a percentage in the Numerical interface, ranging from 0-100% representing low-high confidence.

Figure 2d
Graphical Interface



Note. The classifiers SCI was represented as a bar graph in the Graphical interface, which provided the 5 most probable labels for each image, ordered by their probability.

3.3.3.5 Interface Designs

All four interfaces contained the same basic features but varied in the SCI they displayed (See Figure 2). The Control interface provided no SCI. For interfaces that displayed SCI, this was represented using the model's predicted probabilities (between 0 and 1) for the most likely label(s) to match each image. The Iconography interface provided the simplest form of SCI, discretized as low (<0.33), medium (>0.33 and < 0.66) or high confidence (>0.66), represented by the classifier's label changing color to be red, yellow, or green, respectively. The Numerical condition was more precise, presenting SCI as a text-based numerical percentage, ranging from 0-100% representing low-high confidence. The Graphical condition was the most complex representation of SCI, illustrated as a horizontal bar graph visualizing the distribution of the classifier's 5 most probable labels for each image.

3.3.3.6 Questionnaires

NASA-TLX: After each task block, participants reported their subjective task load when working with each GUI, on a low-high scale (0-100%) (Hart and Staveland, 1988).

Propensity to Trust Machines Questionnaire (PTMQ): A series of 6 questions where participants rated on a 7-point Likert scale how likely they are to trust machines

(Merritt et al., 2013). Half of participants completed the PTMQ before the experiment started, and the rest after completing the experiment.

Debriefing Questionnaire: Participants answered 7 short questions detailing their thoughts about the classifier (Appendix), which they completed following the last block of the experiment. They could also expand on each answer by writing a short paragraph, to explain these thoughts in further detail.

3.3.4 Procedure

All participants read an information sheet explaining the nature of the experiment, before giving written consent. Before the experiment began, they were taught to use the basic elements within the GUI. All participants were briefly informed how AlexNet could provide labels for each image. They were told that in certain blocks AlexNet would also display different forms of SCI, to help support its labelling decisions. They were given further specific instructions about each type of SCI prior to the relevant blocks. In each trial, the participant first rated how familiar they were with the contents of the image. After providing this rating, the classifier then provided the label for each image, to ensure participants' familiarity was not informed by the classifier's label. Participants then rated the accuracy of the classifier's label, and their trust towards the classifier. Lastly, participants decided to keep or replace the classifier's label for the image, before moving to the next trial. Following completion of the experiment and questionnaires all participants were given a debriefing form, which explained the study in further detail.

3.3.5 Analysis

3.3.5.1 *ANOVA*

Our data were not normally distributed, therefore we had to depart from canonical tests and instead opted for a non-parametric alternative: The Aligned Rank Transform ANOVA (ART-ANOVA) (Wobbrock et al., 2011). This test allowed for examination of multiple factors and their interactions within our repeated measures design. Our primary dependent variable of interest was: (1) participants' trust towards the classifier (Trust). In addition to this, we wanted to explore how trust reflected participants' behaviour, and examined (2) how participants decided to accept/reject the classifier's labels for images (Compliance). To assess whether our stimuli selection was balanced (3) we also looked at participants' familiarity with the images presented (Familiarity). Lastly, we considered (4) the average time taken for trials in each combination of conditions, as an objective measure of task load (Trial Time). Consequently, four ART-ANOVA models were conducted, all containing the same three main factors and their interactions: Classifier Performance,

Image Clarity, and Interface, using the ‘ARTool’ package in R version 4.0.2 (Kay & Wobbrock, 2020; R Core Team, 2020). Each ANOVA model contained random slopes to account for multiple observations for each participant, in which they were exposed to each combination of Classifier Performance, Image Clarity, and Interface within the experiment. Additionally, a Kruskal-Wallis test was also conducted to examine the effect of interface on subjective task-load scores (NASA-TLX). Effect sizes were calculated for each main effect using partial eta squared. Pairwise comparisons for significant main effects were carried out using contrasts from the ‘emmeans’ package, with Bonferroni corrections applied to account for multiple comparisons (Lenth, 2020).

I'd expect to see some random slopes, something like `A*B*C + (A*B*C|ID)`, because you have multiple observations for each ID for every level of A*B*C (if I've understood your design correctly).

3.3.5.*2 Additional analyses*

Nonparametric Kendall’s tau correlations were used to examine the relationships between participants’ PTMQ scores and their average trust towards the classifier, as well as their average compliance with the classifier, which we compared across each combination of Classifier Performance and Image Clarity.

3.3.5.*3 Visualisations*

Static and interactive visualisations were created using the ‘ggplot2’ and ‘plotly’ R packages (Wickham, 2016; Sievert, 2020).

3.3.5.4 Data Availability

An anonymised version of this dataset is available through the UK Data Service ReShare repository here: https://dx.doi.org/10.5255/UKDA-SN-854151. The UK Data Service is funded by the Economic and Social Research Council (ESRC) who provided funding for this project.

3.4 Results

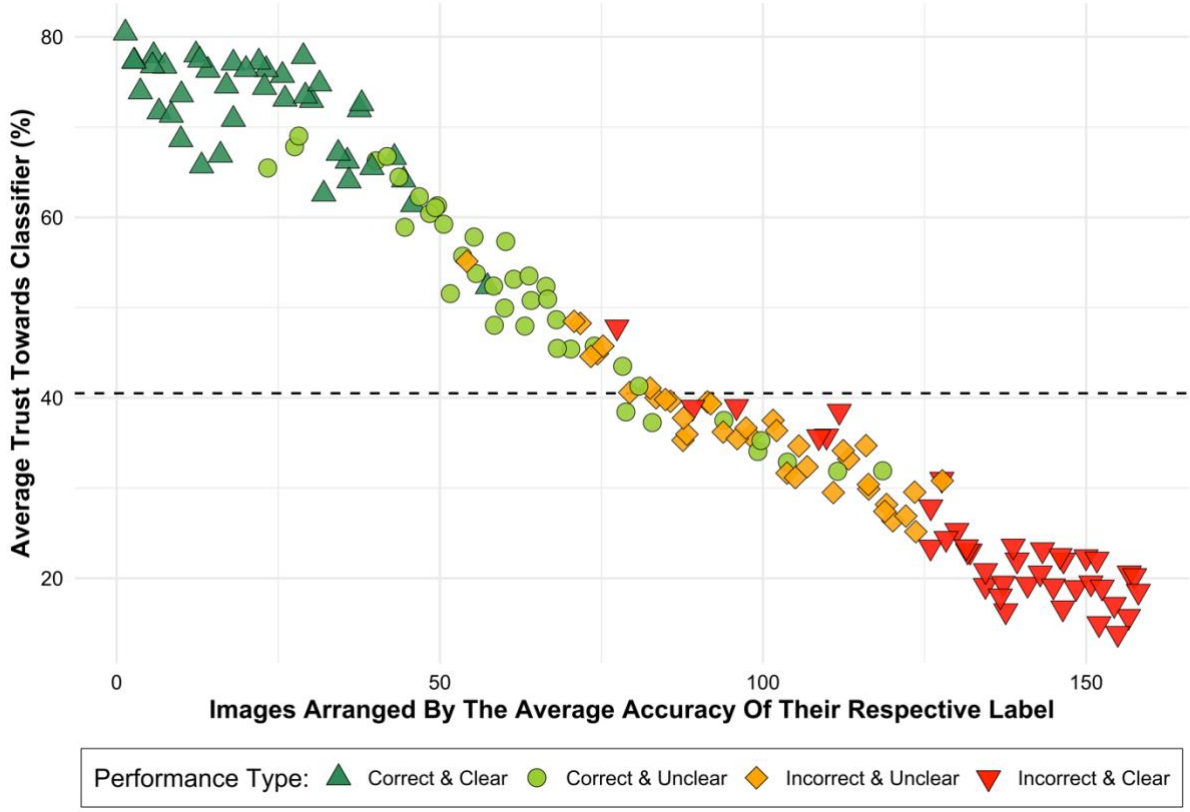
3.4.1 Classifier Performance and Image Clarity

3.4.1.1 *Trust*

Overall, trust was highest in trials where the classifier was correct, and lowest in trials where the classifier was incorrect (Figure 3, and Table 1). However, this relationship was complicated by the clarity of the image. Participants' trust tended to be closest to the grand mean (M=45.77) when processing unclear images, and furthest when processing clear images. For example, if the classifier's label was correct yet the image was unclear (Correct-Unclear: M=51.25, SD=16.02), trust tended to be lower towards the classifier, compared to when the images were clear (Correct-Clear: M=72.07, SD=22.77). Inversely, when the classifier was incorrect trust was higher for unclear images (Incorrect-Unclear: M=36.12, SD=15.50), and lower for clear images (Incorrect-Clear: M=23.62, SD=16.75). ART-ANOVA for Trust revealed a significant interaction between Classifier Performance and Image Clarity $F(1,73)=205.27$, $p<0.001$, $\eta p^2=0.74$, and significant main effects for both Classifier Performance $F(1,73)=226.49$, $p<0.001$, $\eta p^2=0.76$, and Image Clarity $F(1,73)=24.8$, $p<0.001$, $\eta p^2=0.25$. This supports H1a: The classifier's performance was the main driver of trust towards the classifier. This also supports H1b: Image Clarity significantly interacted with system performance when influencing trust towards the classifier.

Figure 3

Trust scores for each image used in the experiment, arranged by accuracy.



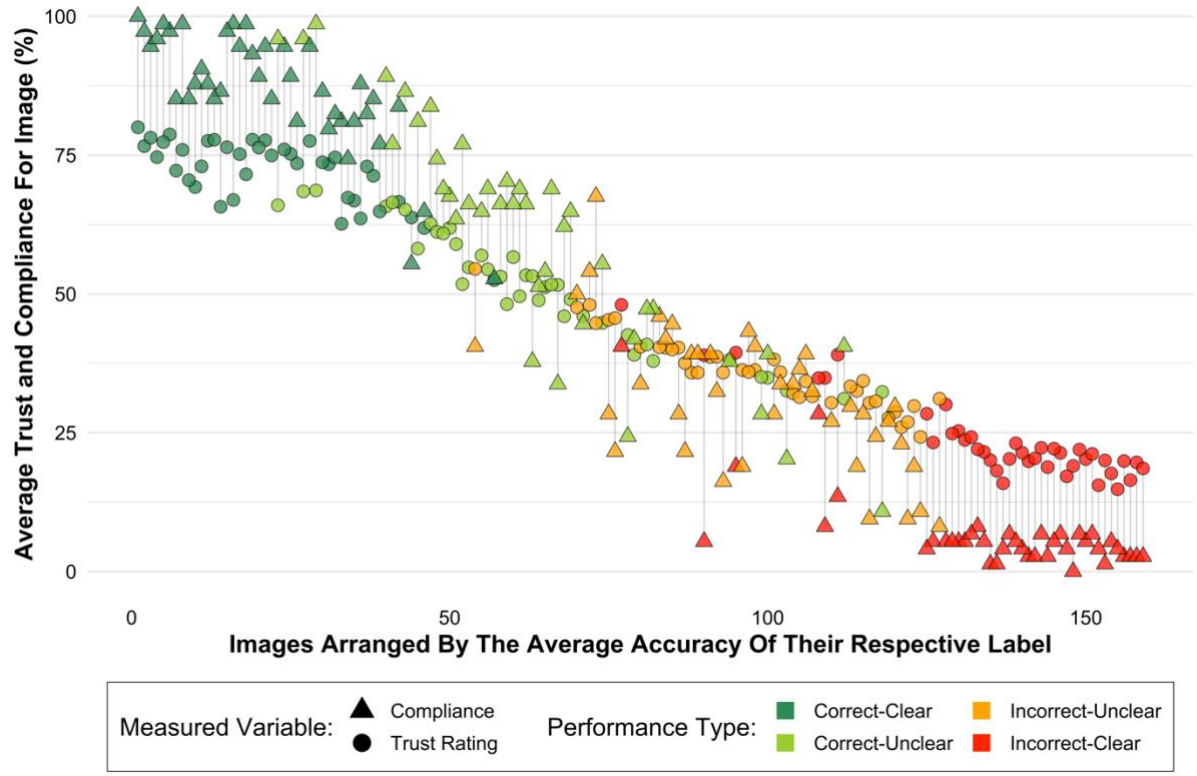
Note. Stimuli arranged by participants' average accuracy rating of classifier's label for the image. Dashed line represents grand median trust.

3.4.1.*2 Compliance*

A similar pattern emerged when examining how participants accepted and rejected the classifier's labels (Figure 4, Table 1). ART-ANOVA for Compliance revealed a significant interaction between Classifier Performance and Image Clarity $F(1,73)=544.24$, $p<0.001$, $\eta p^2=0.88$, and a main effect for Classifier Performance $F(1, 73)=1275.09$, $p<0.001$, $\eta p^2=0.95$. However, there was no significant main effect for Image Clarity $F(1,73)=1.21$, $p=0.27$, $\eta p^2=0.02$.

Figure 4

Difference between trust and compliance for each image used in the experiment, arranged by accuracy.



Note. Stimuli arranged by participants' average accuracy rating of classifier's label for the image.

3.4.1.*3 Familiarity*

In general, participants were more familiar with the images in the Correct-Clear and Incorrect-Clear combinations, and less familiar with the images in the Correct-Unclear and Incorrect-Unclear combinations, as we expected (Table 1). While there was no difference in familiarity between the Correct-Clear (*M*=92.89, *SD*=8.58) and Incorrect-Clear (*M*=92.31, *SD*=9.15) stimuli, there was however a difference between the stimuli in the Correct-Unclear (*M*=41.30, *SD*=14.39) and the Incorrect-Unclear combinations (*M*=29.89, *SD*=12.82). Therefore, we cannot rule out the possibility that some of the differences in Trust and Compliance were related to differences in Image Familiarity in the unclear images. ART-ANOVA for Image Familiarity revealed a significant interaction between Classifier Performance and Image Clarity $F(1,73)=175.22$, $p<0.001$, $\eta p^2=0.71$, and main effects for both Classifier Performance $F(1,73)=226.94$, $p<0.001$, $\eta p^2=0.76$, and Image Clarity $F(1,73)=798.24$, $p<0.001$, $\eta p^2=0.92$.

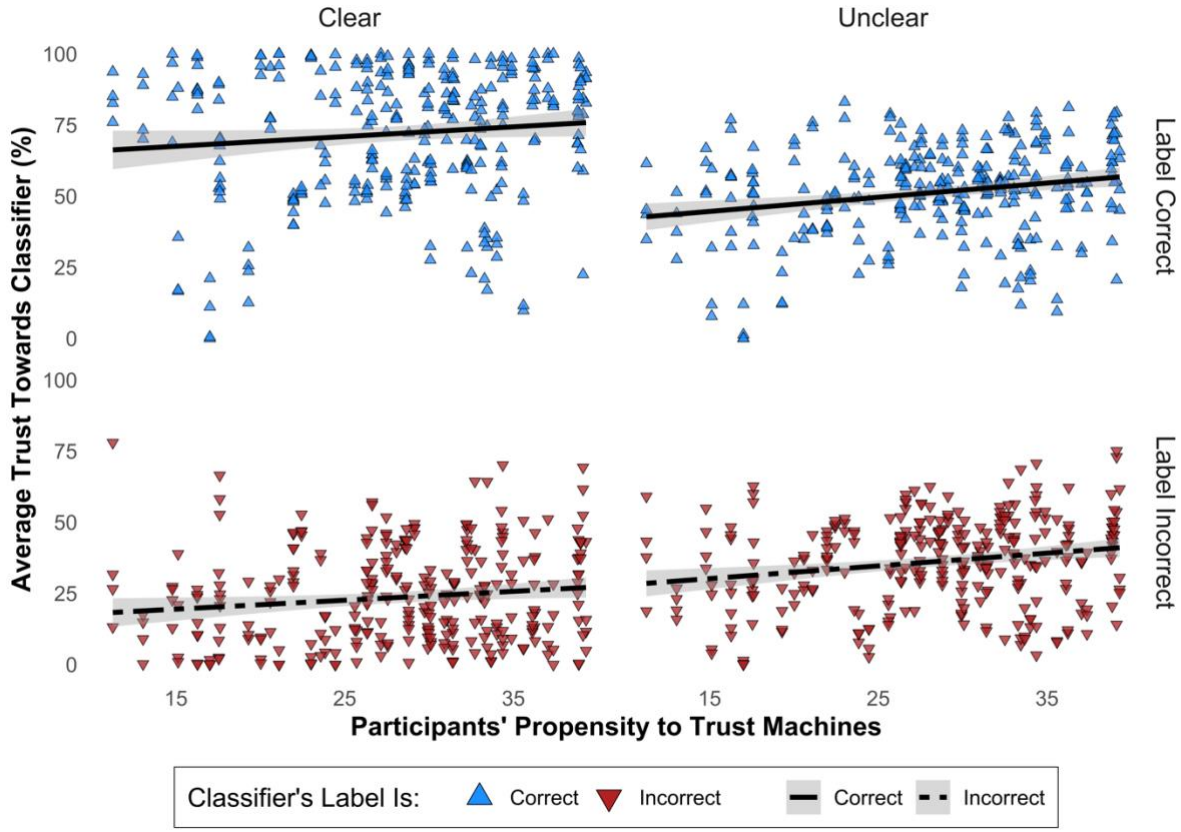
3.4.2 Propensity to Trust Machines

3.4.2.1 Trust

Participants' total scores in the PTMQ were distributed as follows: *M*=28.25, *SD*=7.03, Range=11.25-39.30. PTMQ scores predicted higher trust towards the classifier in three of the four different combinations of Classifier Performance and Image Clarity (Figure 5). While these relationships are relatively weak, they suggest that individual differences may inform trust towards an AICS, particularly when processing unclear images, where system performance may be harder to evaluate. Specifically, participants with higher PTMQ scores were more likely to trust the classifier during Incorrect-Clear trials: r_τ=0.09, p<0.05, Incorrect-Unclear trials: r_τ=0.12, p<0.01, and during Correct-Unclear trials r_τ=0.14, p<0.001, yet interestingly this relationship was not present during Correct-Clear trials r_τ=0.06, p=0.101. Nonetheless, this supports H1c: Participants with higher PTMQ scores tended to trust the classifier more when processing unclear images, where performance may be more difficult to evaluate.

Figure 5

Correlations between participants' PTMQ scores and average trust towards the classifier.

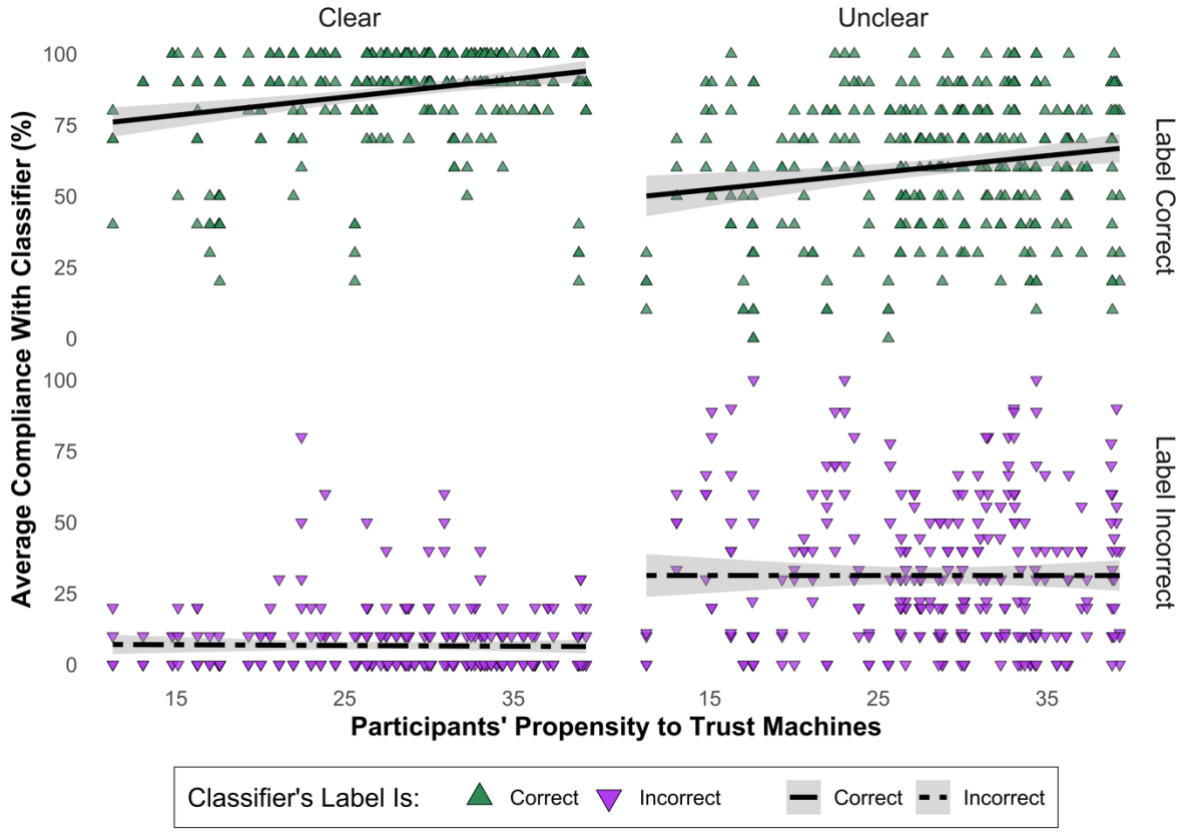


Note. Correlations calculated for each type of performance in each block.

3.4.2.2 Compliance

PTMQ scores predicted higher compliance with the classifier in only 2 of the 4 different combinations of Classifier Performance and Image Clarity (Figure 6). Specifically, participants with higher PTMQ scores were more likely to accept the classifier's label only when the classifier was correct, during Correct-Clear trials: r_τ=0.15, p<0.001, and Correct-Unclear trials: r_τ=0.1, p<0.05. PTMQ scores did not predict greater compliance during Incorrect-Unclear trials: r_τ=0.01, p=0.83, and Incorrect-Clear trials: r_τ=-0.02, p=0.62.

Figure 6
Correlations between participants' PTMQ and average compliance with the classifier.



Note. Correlations calculated for each type of performance in each block.

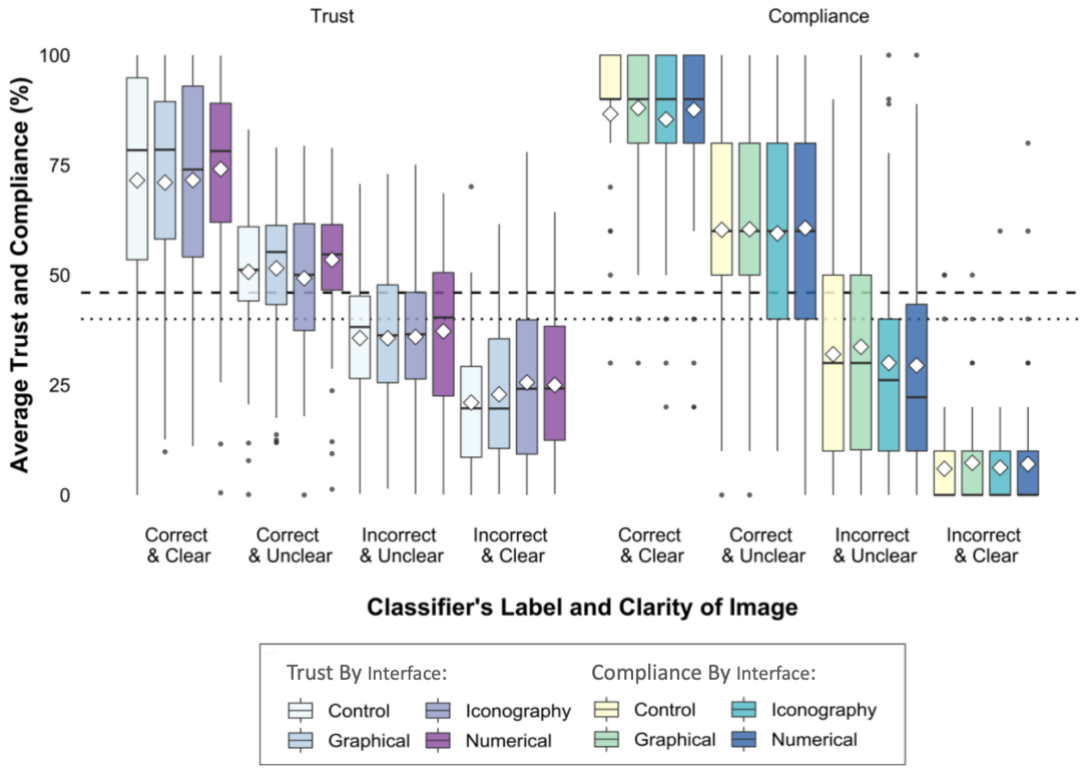
3.4.3 Interface Differences

3.4.3.1 *Trust*

Across the classifier's different interfaces, trust was highest towards the Numerical interface (M=47.45, SD=25.27), and lowest towards the Control interface (M=44.74, SD=26.05) (Figure 7, Table 2). Despite this, trust towards the classifier was not significantly increased when participants worked with the experimental interfaces. While they did not improve trust, most participants reported an explicit preference for working with the interfaces that displayed SCI, suggesting they still found them beneficial on some level (Table 2). ART-ANOVA for Trust revealed no significant main effect of Interface $F(3,219)=1.66$, $p=0.18$, $\eta p^2=0.02$. Thus, H2a was not supported: SCI did not improve overall trust towards the classifier. Moreover, there was no interaction between Interface and Classifier Performance, nor was there between Interface and Image Clarity. Thus, when the classifier's performance was difficult to evaluate, SCI did not improve participants' trust towards the classifier. This means that H2b was also not supported: confidence information did not improve trust when processing unclear images.

Figure 7

Participants' Trust and compliance with the classifier when witnessing each performance type with each interface.



Note. Dashed line represents overall median trust towards the classifier, and dotted line represents median compliance with the classifier. White diamonds represent individual means for each combination. Black dots represent outliers.

3.4.3.2 *Compliance*

Participants were most likely to accept the classifier's label when working with the Graphical interface (M=47.34, SD=35.84), and least likely when working with the Iconography interface (M=45.27, SD=36.17) (Figure 7, Table 2). Despite there being no difference in trust between experimental interfaces, there were small significant differences in compliance with the classifier, suggesting some participants may have been more likely to accept the label provided by the classifier when it provided them with confidence information. ART-ANOVA for Compliance revealed a small main effect of Interface $F(3,219)=3.26$, $p<0.05$, $\eta p^2=0.04$. However, pairwise comparisons suggest the differences between interfaces were nonsignificant: with the most notable being between the Graphical and Iconography interfaces (p=0.098) and between the Graphical and Numerical Interfaces (p=0.134). There were no interactions involving the interface factor.

3.4.3.3 *Post-Hoc Power Analysis for Trust Model*

Post-Hoc power analysis was conducted for examining the influence of SCI interface cues on participants trust scores. Our sample size for this experiment was 74 participants, with 4 observations (averaged values for each combination of classifier performance and image clarity) for each of the 4 interface conditions. The ART-ANOVA model for Trust Scores had a $\eta p^2=0.02$, suggesting we only had a 1-β power of 0.835. To achieve a 1-β power of 0.95 based on a within subjects design of 4 groups with 4 observations, we may have needed a sample size of around 108 for full power in this study.

3.4.5 Task Load

3.4.5.1 *NASA-TLX*

A Kruskal-Wallis test on participant's subjective task load scores revealed no differences between the experimental and control interfaces $H(3)=0.401$, $p=0.94$, (See Table 2). This suggests that the extra information presented by the classifier's experimental interfaces did not increase participants' subjective workload.

3.4.5.*2 Trial Time*

On average, participants spent the most time (seconds) per trial when working with the Graphical interface ($M=19.24$, $SD=7.21$), and the least time with the Control interface ($M=16.43$, $SD=6.83$). This suggests that participants did not necessarily ignore the extra information presented within the experimental interfaces, particularly when working with the Graphical interface. There were also significant differences in Trial Time relating to Classifier Performance and Image Clarity, however these were less interesting. This was because participants were expected to take longer in trials when the classifier was incorrect, given they had to overwrite the classifier's label, and in trials with unclear images, given the classifier's performance is harder to interpret. ART-ANOVA on Trial Time revealed a main effect of Interface $F(3,219)=11.47$, $p<0.001$, $\eta p^2=0.14$, which suggests that participants took longer to complete trials when presented with the classifier's SCI. Pairwise comparisons illustrated most of the significant differences were attributable to the Graphical interface, in comparison to the Control ($p<0.001$), Iconography ($p<0.001$) and Numerical interfaces ($p<0.05$). There were also significant main effects for Classifier Performance $F(1,73)=206.14$, $p<0.001$, $\eta p^2=0.74$, and Image Clarity $F(1,73)=41.77$, $p<0.001$, $\eta p^2=0.36$, as well as an interaction between Classifier Performance and Image Clarity $F(1,73)=91.95$, $p<0.001$, $\eta p^2=0.55$.

Thus, H2c was not completely supported outright, as SCI presented in the experimental interfaces did not increase subjective participants' task load scores. However,

there were significant differences attributable to experimental interfaces when considering the average time spent per trial as an objective measure of task load, with the Graphical interface generally being the most time consuming.

3.5 Discussion

This study sought to understand how individuals calibrated their trust towards an AICS when completing an image classification task. Trust towards the classifier was primarily based on the accuracy of the system's description of images. Trust tended to be highest when the classifier's label was correct, and lowest when incorrect. However, the clarity of the image being processed also influenced trust, such that if the contents of the image were clear then participants were more extreme with their trust, yet with unclear images their trust regressed towards the mean. Moreover, there was also evidence of individual differences amongst participants. The participants with a positive bias towards machines, as indicated by higher scores on the PTMQ, tended to trust the classifier slightly more when processing unclear images. Thus, this study provides an insight into how human users place trust in a system designed to make classifications on image-based data, and expands upon this by also exploring how environmental and interpersonal factors contribute to users' trust towards the system. Additionally, we further built upon this by investigating whether trust towards the classifier could be improved by increasing system transparency through different displays of SCI, yet found little support with the formats we used. The implications of these findings are discussed below.

3.5.1 Trust Towards an AICS

In line with previous research, system performance was the primary driver of trust towards the AICS (Hoff and Bashir, 2015; Yu et al., 2019). This is unsurprising, given autonomous systems are typically designed to handle a specific set of tasks, and therefore task errors represent a violation of their fundamental purpose. However, evaluations of AICS performance seemed to extend beyond simple correct vs. incorrect judgements, as trust towards the classifier varied within correct and incorrect trials. This possibly reflects the nuance in the image classification task, where the classifier must go into more detail than the simpler yes/no type judgements provided by other autonomous systems (Yu et al., 2019; Merritt et al., 2013). At the same time, we should also consider that the classification of images is a relatively familiar task that the human user can often complete by themselves. By contrast in Selkowitz and colleagues' (2017) study, the robotic squad member provided users with various forms of navigational and situational data and is therefore arguably a more complicated task for the user to undertake. Undoubtedly, the greatest benefits of AICS will arise in applied settings when users are tasked with processing large quantities of data, instead of individual images. Nonetheless, our results provide an interesting insight into how individuals perceive the decisions of AICS systems.

For example, when the classifier incorrectly labelled one image of a rowboat as a speedboat participants' average compliance was low, yet trust remained relatively high, despite the error (Figure 4). This illustrates how participants were able to accommodate errors when there is categorical overlap between classifications, and may itself be worth further, more rigorous investigation in future studies.

Evaluations of classifier performance were also informed by how difficult the image was to classify: if the contents of the image were clear, trust was generally higher when correct, and lower when incorrect, compared to when processing images with unclear contents. This appears in line with previous research where trust towards an autonomous baggage scanner was also influenced by the difficulty of the task (Merritt et al., 2013). By building on this, our study illustrates how task difficulty, considered as a component of Situational Trust within Hoff and Bashir's (2015) model, can also influence trust towards AICS. Moreover, participants' compliance with the classifier was also informed by the difficulty of the task. Compliance was typically highest in trials where the classifier was clearly correct, and lowest in trials where it was clearly incorrect. However, this compliance was less uniform in trials with unclear images, suggesting participants were more likely to replace the classifier's labels in difficult trials. Similar to this, changes in task difficulty have been shown to influence how medical practitioners use Clinical Decision Support Systems (CDSS). Goddard and colleagues (2014) report that practitioners were more likely to switch decisions when working with a CDSS in scenarios requiring difficult prescriptions. While this uncertainty may appear detrimental to the operator, Lyell and colleagues (2018) report that using CDSS helped lower users' cognitive load when dealing with more difficult prescriptions. Therefore, when working with autonomous systems to overcome difficult tasks, the advice of the system may still be beneficial even if the system's decision is ultimately replaced or overruled by the operator. While it is worth noting that the increase in difficulty in the previous studies differs from the methods used in the current study, our findings provide further illustration of how changes in the difficulty of the task may influence how operators use autonomous systems.

Additionally, there were individual differences between participants' trust towards the classifier, which may be attributable to their PTMQ scores. Specifically, individuals with higher PTMQ scores tended to have slightly higher trust towards the classifier, particularly during trials with unclear images. Interestingly, higher PTMQ scores also correlated with higher compliance with the classifier, but only in trials where the classifier was correct. This may suggest that while the individuals with higher PTMQ scores tended to trust the classifier more, they remained critical of its performance and their positive bias

did not correspond with higher acceptance of incorrect labels. Within Hoff and Bashir's (2015) model, these individual differences are indicative of Dispositional Trust specific to each operator. The importance of individual differences is also illustrated within models of technology acceptance, which recognize the influence of moderating factors such as the age, gender, and experiences of the operator (Venkatesh et al., 2012). Here, we used convenience sampling in our participant recruitment, and therefore primarily focussed on PTMQ scores as a measure of individual differences. Nonetheless, this echoes previous findings where individuals with higher PTMQ scores and greater Automation Bias tended to place more trust in autonomous technologies (Merritt et al., 2013; Goddard et al., 2014). Therefore, our findings support previous literature suggesting that individual differences can influence trust and attitudes towards autonomous technology. In particular, we demonstrate that biases towards technology could make individuals more likely to trust an AICS when working with it, yet crucially these biases do not automatically translate into making the individual more likely to accept erroneous decisions from the system.

3.5.2 Improving Trust

Both the SAT model (Chen et al., 2014) and Hoff and Bashir's (2015) model suggest that users are more likely to trust autonomous systems with more transparent interfaces. However, we found little support for SCI improving trust towards the AICS, despite previous evidence suggesting confidence information can benefit trust towards autonomous systems (Zhang et al., 2020; Desai et al., 2013). For example, there was no apparent benefit to trust during the more difficult trials with unclear images, despite SCI providing greater information about the classifier's decision. It is possible that the formats we used to convey SCI were not optimal, and that participants were unable to effectively extract the information. This possibility is consistent with previous evidence suggesting that individuals may have difficulty understanding information presented in formats such as bar graphs (Chaphalkar & Wu, 2020; Godau et al., 2016). Likewise, as discussed above, the image classification task itself may have been relatively easy for participants to complete by themselves, meaning that the classifier's decisions, and by extension SCI, may have been of limited use to participants. Additionally, any potential benefits from SCI may have been lost due to the low overall reliability of the classifier within our experiment, which stemmed from our experimental design. Hoff and Bashir (2015) consider system reliability as a subcomponent of system performance, and while design features such as SCI can improve system transparency, any benefits to trust may be lost due to system reliability being more influential than transparency. This could be supported by participants' responses during debriefing, where they rated the classifier as more a tool

than a teammate, and often found it unpredictable (Table 3). Future studies may benefit from employing high and low reliability conditions, in order to explore this further.

Despite this low reliability, participants still found the classifier helpful (Table 3). Moreover, they overwhelmingly preferred working with the classifier's SCI interfaces (Table 2) and did not appear to feel encumbered by the extra information, which still suggests SCI is potentially beneficial. Furthermore, participants spent the most time per trial with the experimental interfaces, particularly the Graphical interface (Table 2). While this does not automatically mean that SCI improved participants' comprehension of the classifier's decisions, it does suggest some processing of this confidence information. While we were primarily interested in trust towards an AICS, it would be beneficial to examine whether SCI can improve users' understanding of these systems. Alongside developing appropriate trust towards autonomous systems, there is also a growing interest in promoting the explainability of autonomous systems, particularly given the 'black box' nature of contemporary machine learning approaches (Abdul et al., 2018). In future studies it would be useful to examine whether displays of SCI can improve the explainability of AICS decisions. This may be particularly well-suited to cases when a classifier assigns the same classification to two distinctly different objects that share similar image features, such as texture and shape.

3.5.3 Beyond Confidence Information

Ultimately, trust towards the AICS could be limited by the way that AICS systems use deep learning techniques when learning to classify images, which can make their decision-making inherently difficult to explain (Gilpin et al, 2019). As a result, these systems may lack the explainability of other autonomous systems, which may make them fundamentally difficult to trust completely (Rudin, 2019). A recent paper by Chen and colleagues (2019) suggested that the decisions of AICS can be made easier to interpret by highlighting important features within sections of an image, through visual cues such as bounding boxes, in order to support the classification for the full image. By doing so, Chen and colleagues (2019) argue that AICS can mimic the reasoning process of humans when classifying images, where the system can illustrate to the user that the classification is based upon shared features with a prototypical image of the classification, essentially: "this image looks like that image". Thus, the 'black box' nature of AICS systems might mean that providing SCI alone could be inappropriate for improving trust, and instead users' trust could ultimately benefit from efforts that make AICS decisions more easily interpreted. Our lack of empirical support for SCI improving trust towards the AICS may be disappointing for potential designers, however these findings still raise important

considerations. Designing interfaces for autonomous systems is a complex process, and based on our evidence, simply providing a single indicator of system decision-making such as SCI, may not be the best way to improve users' trust, at least in the case of AICS. While displays of SCI have shown promise in previous studies, (Verame et al., 2016; Zhang et al., 2020), it may not be a 'magic bullet' for improving trust towards all automation. Nonetheless, while SCI did not explicitly improve participants' trust towards the AICS, most participants still preferred interfaces that displayed SCI, which suggests it might be beneficial to some degree. Thus, this study motivates further research into developing novel methods for conveying the decision making of systems like AICS.

3.5.4 Limitations

This study involved interaction with an AICS in a relatively low stakes task, where participants worked with the classifier to label neutral stimuli. Applied uses of AICS may also include higher stakes tasks, such as identifying patients with diseases (McKinney et al., 2020). In such cases, trust towards an AICS may be even more susceptible to system errors, given the more serious consequences of false alarms and missed cases. By contrast, in our experiment there were no consequences associated with system errors. Regardless, participants still modulated their trust in response to system successes and errors, while they tended to comply with the classifier only when it was correct, suggesting they took the task seriously despite these low stakes. Future studies may wish to build upon these findings by introducing greater consequences for task errors.

3.6 Conclusion

During a human-computer image classification task, trust towards an AICS was primarily based on the classifier's ability to label images. Additionally, image clarity significantly interacted with AICS performance, and further informed participants' ratings of trust and compliance, illustrating the role of task difficulty in their evaluations. Furthermore, some of the variance in trust towards the AICS appeared to have been attributable to individual differences amongst participants, as those with higher propensity to trust machines scores tended to have slightly higher trust towards the classifier. Lastly, while most participants preferred interfaces that displayed system confidence information, it did not appear to improve their trust towards the classifier, despite previous studies suggesting confidence information can improve trust.

3.7 Data Tables

Table 1
Descriptive statistics for Trust Score, Label Accuracy, Image Familiarity, Compliance, and Time as a function of Performance Type.

	Trust (%)		**Label Accuracy (%)**		**Classifier Compliance (%)**		**Familiarity of Images (%)**		**Time Per Trials (Seconds)**	
Performance Type	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Correct & Clear	72.07	22.77	92.19	7.13	86.89	17.77	92.89	8.58	13.10	5.25
Correct & Unclear	51.25	16.02	59.30	12.35	60.20	24.13	41.30	14.39	17.66	7.02
Incorrect & Unclear	36.12	15.50	34.46	13.76	31.29	25.08	29.89	12.82	20.26	8.63
Incorrect & Clear	23.62	16.75	10.32	9.07	6.62	11.35	92.31	9.15	19.98	7.13

Table 2

Descriptive statistics for Trust, Accuracy, Compliance, Familiarity, Time, TLX, Aesthetics and Overall Preference as a function of Interface.

	Trust the Classifier (%)		Label Accuracy (%)		Classifier Compliance (%)		Familiarity of Images (%)		Time Per Trials (Seconds)		Task Load (NASA-TLX)		Aesthetic Rating (1-7)		Favourite Interface
Interface	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*(%)*
Control	44.74	26.05	49.28	32.19	46.21	36.52	64.35	30.73	16.43	6.83	224.20	77.83	4.51	1.43	7
Graphical	45.27	25.44	49.98	31.56	47.34	35.84	63.76	31.40	19.24	7.21	230.68	78.47	4.54	1.39	50
Iconography	45.60	25.25	47.75	32.53	45.27	36.17	63.81	31.44	17.14	8.02	231.43	85.08	4.62	1.40	17
Numerical	47.45	25.27	49.27	32.77	46.18	37.16	64.49	30.59	18.18	8.23	225.57	80.12	4.53	1.48	26

Table 3
Descriptive statistics for responses to Questions 1-6 from Debriefing Questionnaire

	Response (1-7)	
	M	*SD*
How helpful did you think the classifier was? <Not at all / A Great Help>	4.63	0.87
How predictable was the classifier's behaviour? <Predictable / Unpredictable>	4.64	1.38
How specific did you think the classifier's labels were? <Too Specific / Too General>	3.44	1.12
If you had to describe it to someone, how you would characterise the classifier? <Teammate / Tool>	5.48	1.50
If you had to classify another set of images, would you want to work with the classifier again? <With Classifier / Alone>	2.97	1.37
If you had to classify another set of images, which type of collaborator would you prefer? <Computer / Human>	4.54	1.51

Chapter 4: Promoting Better Understanding and Appropriate Trust When Working with an Autonomous Image Classifier

4.1 Abstract

When designing autonomous systems, there are a variety of ways in which the system's decision making can be made more transparent to the user. When provided with relevant decision support information, the user may trust the system more appropriately, and may also report better understanding of the system's decisions. We investigated 2 different cues of decision support information (System Confidence Information and Gradient-weighted Class Activation Mapping) to see whether they can promote trust towards an autonomous image classifier. We also explored whether these cues improve users' self-reported understanding of the system's decision making. These cues were provided intermittently across 8 experiment blocks, in which participants collaborated with the classifier to complete an image classification task featuring 240 single-image trials. Participants worked with each of the 4 interfaces twice, across both conditions of High Reliability (90% correct) and Low Reliability (60% correct). Results suggest that trust was not improved by the addition of any of the experimental cues. However, participants reported greater understanding of system decision making when working with the interface that provided Gradient-weighted Class Activation Mapping. This suggests that users may benefit from cues that visualise the image areas/features that are relevant to classifier decision making. Additionally, participants typically preferred the classifier interface that provided the most detailed version of the systems' decision support information, suggesting these cues remain useful to users.

4.2 Introduction

For all the ingenuity and innovation required to develop new technologies, their ultimate success depends on their adoption by human users. Originally, models of technology adoption centred only on the perceived usefulness and ease of use of the system (Davis, Bagozzi, and Warshaw, 1989). However, extensive work by Viswanath Venkatesh has since highlighted many other potential factors which can inform a human user's adoption of technologies, including habits, their previous experiences, and even the price of the system (Venkatesh & Davis, 2000; Venkatesh et al., 2012; Venkatesh, 2015). However, for autonomous technologies to be successfully adopted, there is an additional emphasis on the user being able to trust the system, given it will be operating with a degree of independence from the user.

Blindly trusting an autonomous system is inappropriate; instead, trust should be calibrated to ensure that it accurately reflects the performance of the autonomous system (Muir, 1987). In doing so, the user will not place too much trust in an error-prone machine (mistrust), nor will they place too little trust in a competent system (distrust) (Parasuraman & Riley, 1997). However, evidence suggests that there are many kinds of factors which can influence users' trust towards autonomous technologies. Some are relatively straightforward: system performance and task accuracy closely align with trust towards autonomous systems (Yu et al., 2019; Papenmeier et al., 2019). Some are less immediately apparent, and illustrate more complex psychological influences on users' trust: individuals working with cognitive agents preferred systems with more anthropomorphic features i.e. human-like traits (De Visser et al., 2016). Hoff and Bashir (2015) integrate these different factors by considering 3 layers of trust that inform the way individuals use autonomous systems. Dispositional trust covers human-centric factors related to the user, situational trust covers operational task and environment-related factors, whilst dynamic trust covers factors related to the autonomous system itself, and the user's experience of similar systems (Hoff & Bashir, 2015). Ultimately, these layers are combined to inform the user's continued reliance on the system. Therefore, the successful implementation of autonomous technologies is closely linked to their ability to garner appropriate trust from users.

4.2.1 Trust Towards Autonomous Image Classifiers

In our previous experiment, we explored how human users calibrated their trust specifically towards an autonomous image classifier system (AICS), when collaborating in an image classification task (Ingram et al., 2021). AICS are autonomous systems which can be trained to process large quantities of image data, in order to reduce the demand on human collaborators. Using Hoff and Bashir's (2015) model, we examined how dispositional, situational, and system-based factors all influenced users' trust towards the AICS. Our experiment was separated into two main components: 1. Examining how participants' trust was linked to the performance of the classifier, and 2. Exploring whether trust towards the classifier could be improved by providing additional information about system performance.

From the first component, we found that users' trust was primarily based upon the classifier's performance; whether it was able to correctly identify the contents of the image. When the classifier was correct trust towards the classifier was typically higher, whilst trust tended to be lower when the classifier was incorrect. This relationship between trust and system performance was also influenced by the quality of the image being processed. When the contents were clear and easily identifiable, participants were more

decisive with their trust ratings; high trust when correct, and low trust when incorrect. However, when processing low quality images (i.e. blurred and partially occluded contents), participants were less decisive with their trust, and often gravitated towards the midpoint of the scale we provided when reporting their trust. This suggested participants' evaluations of the AICS were also informed by the difficulty of the task. There was also evidence of individual differences in trust towards the AICS; participants' scores in the Propensity to Trust Machines Questionnaire (PtTMQ) (Merrit, 2011) correlated positively with increased trust towards the classifier. These findings were in line with Hoff and Bashir's (2015) model and also replicated findings from a similar study involving an automated baggage scanner (Merritt et al., 2013). However, our findings were less straightforward within the second component of our experiment, when we attempted to improve participants' trust towards the classifier by displaying the system's decision-making information.

4.2.2 Improving Trust Towards an AICS

Evidence suggests that providing users with more detailed information about system processes can increase trust towards the system, as it can make the system's decision-making more transparent (Chen et al., 2014; Desai et al., 2013; Mercado et al., 2016; Selkowitz et al., 2017; Tomsett et al., 2020; Zhang et al., 2020). A popular method for improving transparency is the use of System Confidence Information (SCI), which illustrates the system's uncertainty about its decisions, and has previously improved users' trust towards a variety of autonomous systems (Desai et al, 2013; Verame et al., 2016; Zhang et al., 2020). In Ingram and colleagues (2021) we compared three different forms of SCI by presenting the classifier's confidence in the following ways: 1. Confidence discretised with colours (Iconography), 2. Confidence as a percentage (Numerical), 3. Confidence as a bar graph visualising the 5 most probable labels for each image (Graphical). Despite previous findings, we found limited support for SCI improving participants trust towards the AICS. When compared against a control interface without SCI, there was no obvious increase in trust towards the three experimental interfaces featuring SCI. However, when we asked participants to state their explicit preference for one of the four interfaces used in our experiment, they overwhelmingly preferred the interfaces featuring SCI. Moreover, the strongest preferences were for the Graphical and Numerical interfaces, which provided the most detailed form of SCI. Thus, while SCI did not improve trust towards the classifier, it appears participants may have found the SCI beneficial in some capacity. We speculated that this may be explained by the relatively low reliability of the AICS throughout the experiment (discussed in greater detail later).

The other factor that we have since considered was whether the nature of the image classification task may have meant that SCI was less useful to users, as it only illustrated the system's certainty when classifying images. It is possible that trust towards the AICS could be improved by presenting users with a more meaningful cue of system decision-making. It is also possible that SCI may not have been easily understood by some participants, as it may require some understanding of how probabilities work in order to be useful. Thus, instead of comparing different forms of SCI against each other, in the current research we compare one form of SCI (an amalgamation of the Numerical and Graphical interfaces used in Ingram et al., 2021) with an entirely different cue of classifier decision-making: Gradient-weighted Class Activation Mapping (commonly known as Grad-CAM or CAM).

4.2.3 Gradient-weighted Class Activation Mapping

CAM is method for visualising the activity of neural networks, and can be used to highlight the discriminative image areas used by an AICS when identifying the contents of image data (Jia & Shen, 2017; Selvaraju et al., 2017; Yang et al., 2019; Mukhopadhyay et al., 2020). Simply put, this allows for visualisation of the areas which are most influential for each classification, often illustrated in the form of a heatmap where warmer colours represent greater activity (Selvaraju et al., 2017; Yang et al., 2019). For example, Jia and Shen (2017) employed CAM to illustrate how their neural network identified skin lesions, by visually demonstrating that the majority of classifier activity was centred on areas of damaged tissue. In theory CAM represents an intuitive way to illustrate the decision-making of an AICS to human users. Chen and colleagues (2019) suggest that cues such as CAM can be used to make the decision making of AICS more interpretable, by mimicking the way that humans explain their decisions when making classifications. However, Rudin (2019) cautions that CAM may not be the perfect way to support AICS decisions. She suggests that even if a user knows there is greater network activity around certain parts of an image, this does not necessarily translate into the user knowing what the network actually does with those areas. Nonetheless, given increased transparency can improve trust towards other autonomous systems, we believe that it remains equally possible that both SCI and CAM may be beneficial to users' trust. While SCI illustrates the classifier's certainty for the label(s) it provides, CAM illustrates the areas of the image that were most informative in the classification, and it is unclear whether one cue will be more beneficial than the other for improving users' trust towards an AICS. Previous studies have typically focussed on investigating ways to increase the spatial accuracy of CAM methods (Patro et al., 2019; Yang et al., 2019), yet to our knowledge there have been no formal

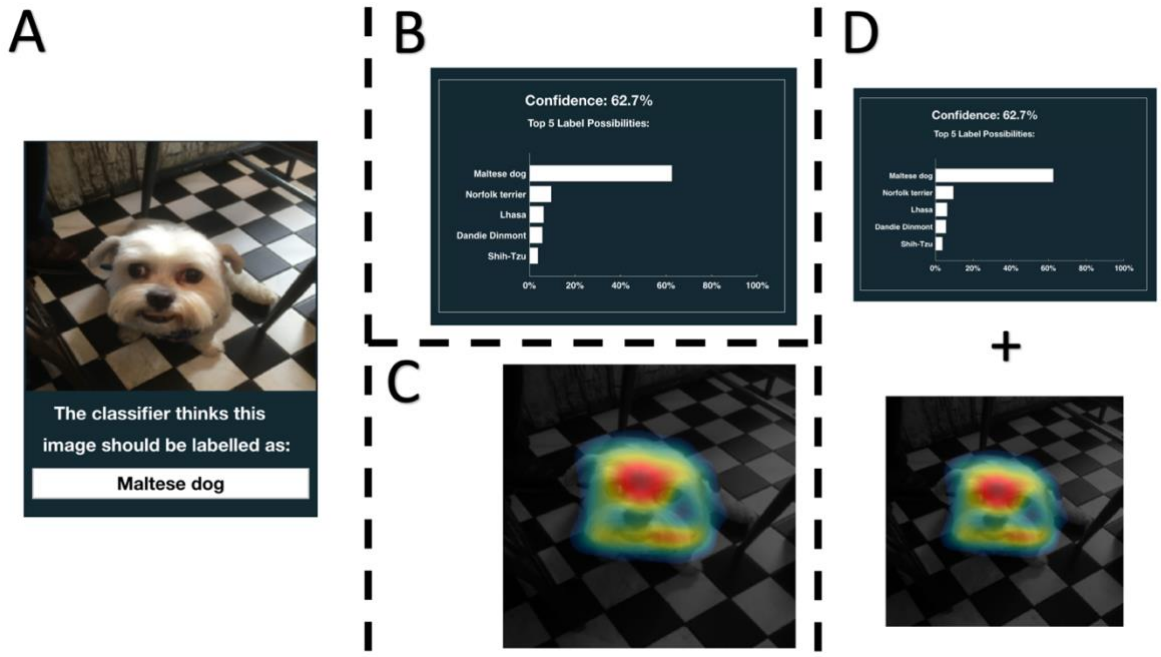
investigations of how CAM can be used to inform trust towards an AICS. Thus, we used an additive model to assess whether trust towards an AICS can be improved by making the system more transparent through both CAM and SCI. We created 3 experimental interfaces which provided different cues to support AICS classifications, which were compared to a basic interface with no cues of decision-making (Control - See Figure 8A). These experimental interfaces were: 1. System Confidence Information only (SCI - See Figure 8B). 2. Class Activation Mapping only (CAM - See Figure 8C). 3. Confidence Information and Class Activation Mapping (SCI+CAM - See Figure 8D). By providing participants with these different interfaces, we were able to examine the utility of both SCI and CAM cues and hypothesised:

(H1a): Trust towards the classifier will be increased when working with the experimental interfaces.

(H1b): Trust will be highest when working with the SCI+CAM interface, as it provides the most detailed information.

Figure 8

Decision Support Information offered by the Classifier within each Interface.



Note. Control Interface (A) only provided labels for images, while additional information was provided in SCI (B), CAM (C) and SCI+CAM (D) interfaces. Further details are provided within the Methods section.

4.2.4 Understanding the Classifier

As autonomous technologies have become more capable of undertaking complex tasks, there has also been an increasing interest in promoting the explainability of these systems, particularly with neural networks which are often criticised for being 'black box' systems (Abdul et al., 2018; Ribera & Lapedriza, 2019). Explainability is characterised as the system's ability to convey the reasoning behind its decision-making, thereby facilitating greater understanding from human users (Gilpin et al., 2019). Theoretically, there may be some overlap between a user's understanding of an autonomous system and their subsequent trust towards the system. If the user is more capable of understanding the decision making of an autonomous system, then it is plausible they may be better placed to calibrate their trust towards the system more appropriately. That being said, trust towards automation should not be entirely dependent upon the user's understanding of how the system works, particularly with more complex systems. For example, an operator may be able to appropriately trust an autonomous system designed to manage a swarm of drones, even though they may never be able to fully understand the how the system coordinates so many different streams of information. Ultimately, comprehension of the decision-making of autonomous systems will undoubtedly vary between different machines and tasks.

When evaluating the performance of an AICS, there is significant scope for human users to understand the decision-making of the system, given visual identification and

object classifications are inherently familiar tasks for humans (Doniger et al., 2001). In Ingram and colleagues (2021), SCI did not improve trust towards the AICS, however we also speculated that most participants may still have preferred the interfaces with SCI because they offered better insight into the classifier's decision-making. While we anticipate that SCI and CAM may improve user's trust towards an AICS, we also expect that these cues may increase the explainability of the system, and thereby improve users' understanding of the system's decision-making. However, Rudin (2019) argues there should be distinction between an explainable system and an interpretable system, wherein an explainable system is retroactively explained to the user through cues such as CAM or SCI, while an interpretable system would be inherently understandable by the user. Rudin (2019) also argues that developers should put more emphasis on designing systems which are interpretable, particularly when the system is intended to make high-stakes predictions, such as in parole decisions. While these are important points that should be addressed in future work, the design of interpretable systems is outwith the scope of this research, and we will primarily focus on how CAM and SCI can increase the explainability of AICS decision-making. We anticipated that both CAM and SCI would improve comprehension of classifier decisions, given both cues increase system transparency, relative to the control interface. Moreover, in line with the Chen and colleagues (2014) and Selkowitz and colleagues (2017), we hypothesised that the participants' understanding of classifier decisions would be highest when provided with both SCI and CAM cues together:
(H2a): Participants' understanding of the classifier will be increased when working with the experimental interfaces.
(H2b): Participants' understanding will be highest when working with the SCI+CAM interface, as it provides the most detailed information.

4.2.5 System Reliability and Trust

In Ingram and colleagues (2021) we suggested our results may have been tempered by the low overall reliability of the classifier throughout the experiment (50%), which stemmed from our experimental design. From this we speculated that any potential benefits to trust from displaying SCI may have been lost, due to the low overall reliability of the system. Indeed, a similar study reported that users' trust towards an automated text classifier appeared to be based more on the system's accuracy, rather than the system's explanations for its decisions (Papenmeier et al., 2019). Likewise, Wright and colleagues (2019) suggest that system reliability had a profound influence on users' trust towards an autonomous military squad member, with system transparency having much less influence. This suggests that the potential benefits of increased system transparency may be

dependent on the overall reliability of the system. In our previous experiment, our analysis was primarily focussed on the performance of the AICS in the individual trials within our experiment, yet we did not manipulate the system's overall reliability across the experiment. However, evidence suggests that the reliability of the system can have a significant influence on users' trust towards autonomous systems (Sauer et al., 2016; Sauer & Chavaillaz, 2017; Hussein et al., 2019). For example, when supervising an autonomous squad member, individuals reported higher trust in settings where the system was perfectly reliable (100%), compared to when the system was relatively unreliable (67%) (Wright et al., 2019). Therefore, we built upon the work of Ingram and colleagues (2021) by examining trust towards an AICS in conditions of high reliability (90% Correct trials) and low reliability (60% Correct trials). We hypothesised that participants' trust will likely reflect the frequency of errors committed by the AICS within each block:
(H3): Trust towards the classifier will be highest in high reliability conditions.

4.2.6 Operators Workload

It is also important to note that with certain autonomous systems, the way that the operator uses the system can directly impact upon the system's reliability (Ozdemir & Kumral, 2019). Therefore, if the human user does not use the system properly, this could lead to sub-optimal system performance, which could potentially create a negative feedback loop influencing users' trust towards the system. This is particularly important given errors and poor system performance can often require users to correct the system's mistakes, thereby increasing the human operator's workload (Sauer et al., 2016; de Visser & Parasuraman, 2011). As such, we expected to see a significant increase in participants' workload when the AICS presents low reliability, given participants will be required to correct the classifier more frequently. We collected subjective workload ratings through the NASA Task Load Index (NASA-TLX), and hypothesised:

(H4a): Participants' subjective task load scores will increase during low reliability conditions.

While we believed that both the SCI and CAM cues have the potential to be equally informative to users, we anticipated that participants may find the CAM interface to be slightly more user-friendly as it arguably requires less effort to interpret. That being said, when we provided cues of SCI to participants in Ingram and colleagues (2021), we did not see an increase in participants' subjective task load. This suggests that providing a single cue of either CAM or SCI alone may not result in an increase in users' subjective task load. However, in our previous study we did see a significant increase in the average time spent per trial when participants worked with the most detailed version of the SCI interface. This

suggests that displaying cues such as SCI and CAM can indirectly increase the user's workload by giving them more information to process. However, the user may still not report a subjective increase in their workload if they find the additional cue(s) to be beneficial. Indeed, increasing an autonomous system's transparency through displays of more information may not lead to higher greater workload in operators if the interface is well designed (Mercado et al., 2016; Selkowitz et al., 2017). Nonetheless, it remains possible that providing both SCI and CAM together at the same time may significantly increase users' workload, as users will be provided with 2 additional cues of system decision-making to process. To assess whether participants become encumbered by SCI and CAM cues, we examined participants' subjective workload (NASA-TLX), as well as the average time spent per trial when working with each interface, and hypothesised:

(H4b): Participants' task load will be significantly higher when working with the SCI+CAM interface.

4.2.7 This Study

Thus, in this study we attempted to build upon the study protocol used by Ingram and colleagues (2021), and address questions raised by our findings. We have three main avenues of interest, which are reflected in our experimental design: (1.) Can trust towards the classifier be improved through greater transparency using SCI and/or CAM cues? (2.) Does this transparency improve individuals' subjective understanding of the classifier's decision-making? (3.) How does the reliability of the classifier inform user's trust towards an AICS? Additionally, we also assessed (4.) How system reliability and interface transparency impacted upon the workload of human users when collaborating with the classifier.

4.3 Methods

4.3.1 Participants

A total of 49 participants (32F), primarily university students (Mean Age = 26.3, Min = 19, Max = 50), were recruited through the University of Glasgow's School of Psychology subject pool. All participants were compensated at a rate of £6 per hour for their time. 50% of participants considered themselves native English speakers. Ethical approval was obtained from the University of Glasgow's College of Science and Engineering ethics committee. Data collection was paused during the COVID-19 pandemic, and was resumed when restrictions were lifted to allow for face-to-face data collection.

4.3.2 Design

A 4x2x2 within-subjects design was used, where participants used each of the 4 Classifier Interfaces (Control, SCI, CAM, SCI+CAM) within each level of Block Reliability (High Reliability, Low Reliability) (8 Blocks total). In each single-image trial (8x30, n=240), the classifier's label would either correctly or incorrectly match the image displayed, characterised as Trial Performance (Correct, Incorrect). The proportion of these correct and incorrect trials defined the reliability for each of the 8 blocks (4x High Reliability = 90% Correct, 4x Low Reliability = 60% Correct). The ordering of blocks was pseudo-randomised with the experiment effectively split into two halves. In blocks 1-4, participants worked with each interface once, with 2 blocks randomly selected to be high reliability, and 2 blocks that were low reliability, and were randomly ordered. In blocks 5-8 participants worked with each interface again, with the alternative reliability for each interface, which was again randomly ordered. The ordering of trials was fixed within each block, and defined by the reliability condition (Figure 9). The average participant took 21.5 seconds to complete each trial, and 10.7 minutes per block.

4.3.3 Materials

4.3.3.1 Image Classifier

Participants interacted with an AICS based on the SqueezeNet image classifier model (Iandola et al., 2016), which used MATLAB's Deep Learning and Image Processing Toolboxes (MATLAB ver. R2019a). SqueezeNet is a pretrained convolutional neural network, trained to classify objects within a 227x227-pixel net. To process each image, the file must first be resized to fit these dimensions, after which SqueezeNet can interpret the image. Like AlexNet, SqueezeNet can provide the probabilities for multiple possible labels for each image, which allowed for creation of the SCI cue, to illustrate the classifier's other possible labels for each image. Crucially, SqueezeNet also allows for representation of the CAM, which can illustrate the classifier's activity in a heatmap-style cue to participants. While it is also possible to implement CAM with models such as GoogleNet and ResNet-18, SqueezeNet provides CAM with a higher spatial resolution than other models (Mathworks, 2019). This effectively means the generated heatmaps bind more tightly to the objects in the images and should therefore be more informative to participants.

4.3.3.2 Images

240 images were selected from the Open Images Dataset V5 (OIDV5) (Now: V6+) (Kuznetsova et al., 2020), none of which were featured in our previous experiment (Ingram

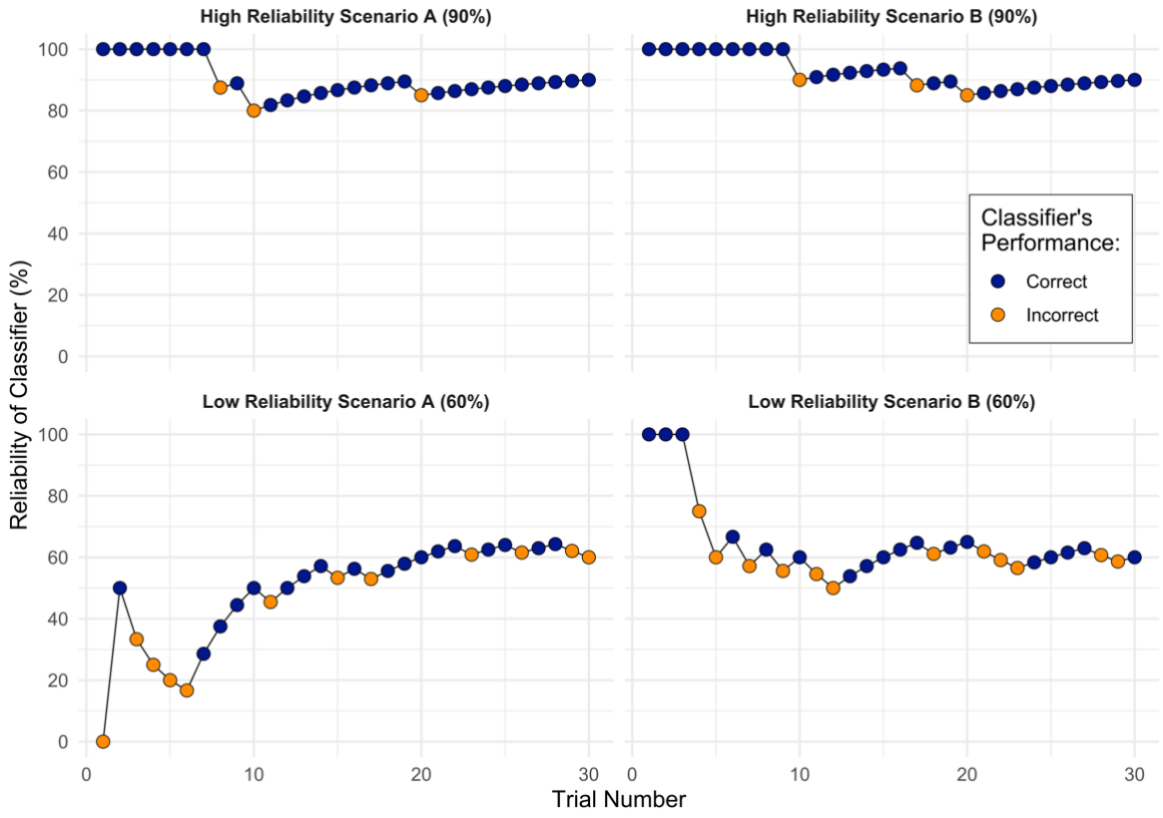
et al., 2021). These were used to create eight separate sets of 30 single-image trials, which featured similar content such as household objects, food items, vehicles, and animals. The classifier's performance was considered as correct when SqueezeNet provided labels which appropriately matched the image's original label in OIDV5, otherwise performance was considered incorrect. Trial performance was intrinsic to each image; performance only varied between images, and participants only saw each image once.

4.3.3.3 Reliability

The classifier's reliability was blocked into 2 different levels: High Reliability (90% of trials correct), Low Reliability (60% of trials correct), which was achieved by controlling the proportion of correct and incorrect trials within each of the 8 sets of images. In high reliability image sets there were 27 correct trials and 3 incorrect trials (90%), with 18 correct trials and 12 incorrect trials in low reliability sets (60%). In each block the location of correct and incorrect trials was fixed, and was pre-defined based on the reliability level. This was done to ensure participants experienced similar types of performance with each classifier interface for each level of reliability. However, to ensure that this did not make the classifier's performance become too predictable, we created 2 scenarios per reliability level (Figure 9). Therefore, each participant completed 2 blocks for each of the 4 reliability scenarios. While the location of correct and incorrect trials was fixed within these scenarios, the images used for correct and incorrect trials was randomly drawn from each set to reduce order effects. This ensured different participants were less likely to see the same images during the same trials.

Figure 9

Classifier's reliability in each of the 4 reliability scenarios.



4.3.3.4 Image Classification Task

The experimental task followed a similar protocol to the one used by Ingram and colleagues (2021). Participants used a mouse and keyboard to interact with the classifier's Graphical User Interface (GUI), which was built using MATLAB App Designer (MATLAB ver. R2019a). In each trial, the target image appeared in the centre of the GUI, and the classifier's label appeared below. Diverging from Ingram and colleagues (2021), we simplified how participants chose to accept or reject the classifier's labels. At the end of each trial, participants used one of three buttons to make the appropriate decision for the classifier's label. If they believed it was correct, they used the green 'Keep' button to move to the next trial. If they believed it was incorrect, they used the red 'Edit' button to manually overwrite the classifier's label with their own, before pressing the green 'Submit' button to move to the next trial. If they had difficulty deciding what the contents of the image were, or did not understand the classifier's label, they used the yellow 'Unsure' button to signify this, which also moved them to the next trial. Additionally, in Ingram and colleagues (2021) we asked participants about their familiarity with the contents of each image, however some participants found this wording confusing. Therefore, we attempted to simplify this by asking participants how easy they believed it was to label the image themselves in each trial instead.

Like Ingram and colleagues (2021), participants rated the classifier's performance on a visual analogue scale within the GUI, using 4 different interactive sliders. These

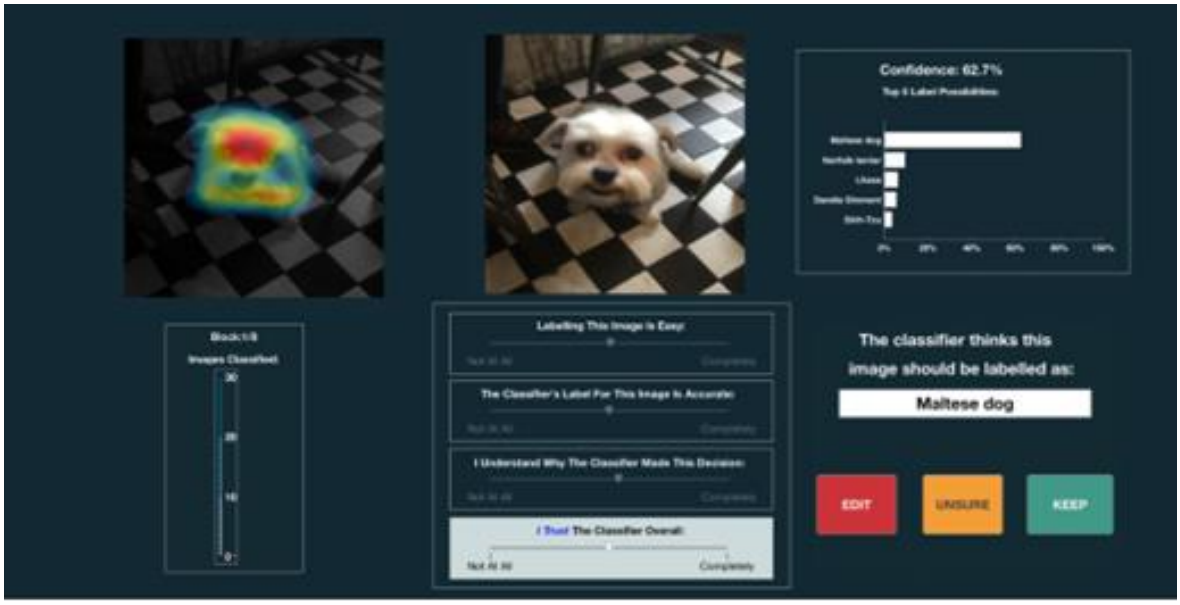
corresponded with: 1.) How easy it was to label each image, 2.) How accurately they believed the classifier's label described each image, 3.) Whether they understood why the classifier made each decision, and 4.) Their overall trust towards the classifier. They were instructed that ratings of label accuracy should reflect the classifier's performance in each individual single image trial, while ratings of trust should represent their continuous interaction with the classifier throughout the experiment. When reporting their understanding, participants were asked to illustrate if the classifier's decisions made sense, irrespective of whether it was correct or incorrect, based on the features within each image. All sliders went from 0-100%, represented with visual anchor points of "Not at all" and "Entirely". Data were collected from each slider after each trial and would reset to the midpoint (50%) between trials. Each slider would change colour (white) to cue participants towards which rating to provide next, guiding the participant throughout each trial. Compliance with the classifier was defined as a trial in which the participant did not use the 'Edit' button.

4.3.3.5 Interface Differences

All four interfaces contained the same basic features, such as buttons, sliders, and a progress gauge which illustrated the number of images processed by the participant (Figure 10). The control interface was the default interface and featured no additional cues of classifier decision-making (Figures 8 & 10). The SCI interface added a display of the classifier's confidence for each image, which was illustrated as a percentage for the top label choice, as well as a bar graph underneath illustrating the distribution of confidence for the 5 most likely labels for the image (Figures 8, 10 & 11). The SCI interface combined features from the two most popular formats of SCI used in used in Ingram and colleagues (2021), which were the bar graph and the numerical percentage. The CAM interface added a display of the classifier's activation map for each image, which was illustrated as a heatmap-style cue overlaid over the image, to illustrate areas of the image with the most activations (Figures 8, 10 & 12). The SCI+CAM interface added both the SCI and CAM cues, to provide participants with the most information each trial (Figures 8 & 10).

Figure 10

SCI + CAM Interface demonstrating classifier GUI, with added CAM and SCI cues.



Note. Control interface features neither CAM or SCI cues.

Figure 11
Classifier Confidence Information (SCI) illustrating the most probable labels.

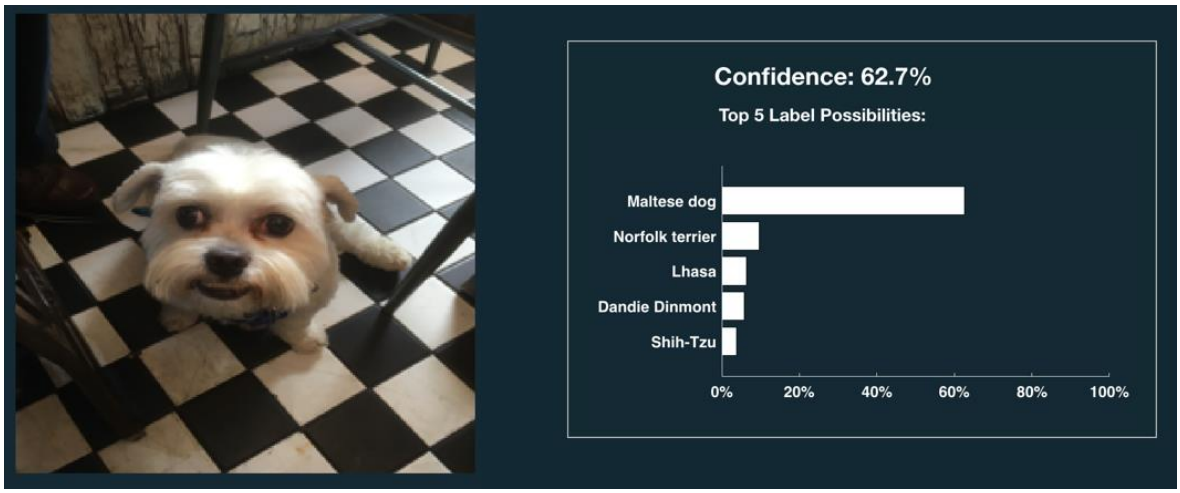
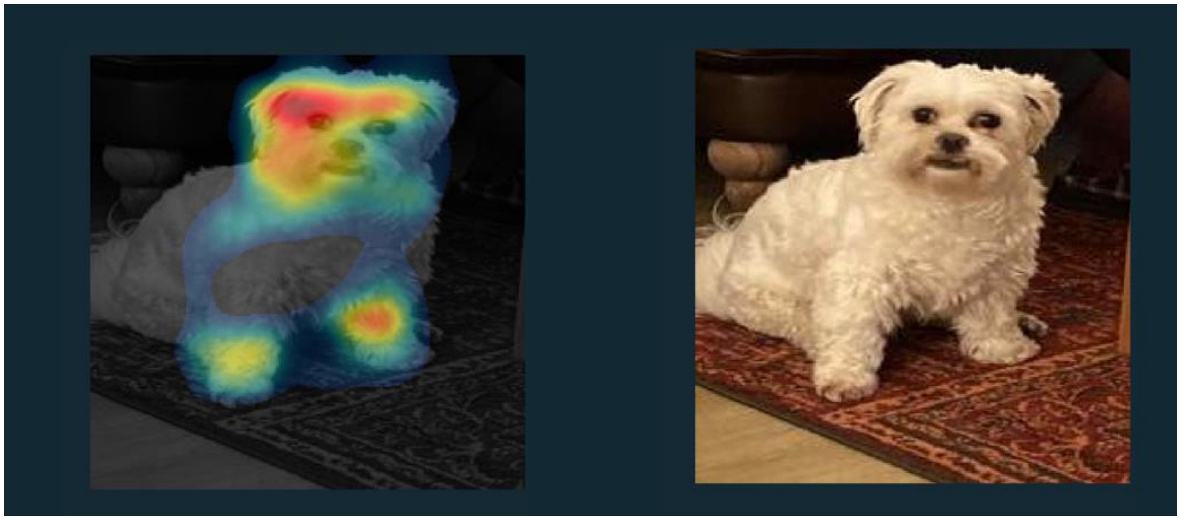


Figure 12
Class Activation Map (CAM) showing areas of image with most activity.



Note. Warmer colours illustrate greater network activity when processing the image.

4.3.3.6 Questionnaires

NASA-TLX: After each task block, participants reported their subjective task load when working with each GUI, on a low-high scale (0-100%) (Hart and Staveland, 1988).

Propensity to Trust Machines Questionnaire (PTMQ): A series of 6 questions where participants rated on a 7-point Likert scale how likely they are to trust machines (Merritt et al., 2013).

Debriefing Questionnaire: Participants answered 9 short questions detailing their general thoughts about the classifier (Appendix). They could also expand on each answer by writing a short paragraph, to explain these thoughts in further detail.

4.3.4 *Procedure*

All participants read an information sheet explaining the nature of the experiment, before giving written consent. The participants then completed the PTMQ. Before the experiment began, they were taught to use the basic elements within the GUI. All participants were briefly informed how SqueezeNet could provide labels for each image. They were told that in certain blocks SqueezeNet would also support its decisions with different cues, which were explained in further detail prior to the relevant blocks. In each trial, the participant first rated how easy they thought it would be to label the image. The classifier then provided the label for each image, to ensure participants' ability to label the image was not informed by the classifier's label. Participants then rated the accuracy of the classifier's label, and then rated whether they understood why the classifier made the decision. Lastly, they rated their overall trust towards the classifier. After this, participants decided to keep or replace the classifier's label for the image, or to report if they were unable to evaluate the label. After each block, participants completed the NASA-TLX. Following the experiment, all participants completed a short debriefing questionnaire, to

give their perceptions of the classifier, before being given a debriefing form, explaining the study in further detail.

4.3.5 Analysis

4.3.5.1 ANOVA

We performed non-parametric analyses using the Aligned Rank Transform ANOVA (ART-ANOVA) (Wobbrock et al., 2011). This test allowed for examination of multiple factors and their interactions within our repeated measures design. Our primary dependent variable of interest was: (1) participants' trust towards the classifier (Trust). In addition to this, we wanted to explore (2) how well participants understood the classifier's decisions (Understanding). We also considered (3) the average time taken in trials within each combination of conditions (Trial Time). Consequently, three primary ART-ANOVA models were conducted, all containing the same three main factors and their interactions: Interface, Block Reliability, and Trial Performance, using the 'ARTool' package in R version 4.0.2 (Kay & Wobbrock, 2020; R Core Team, 2020). Each ANOVA model contained random slopes to account for multiple observations for each participant, in which they were exposed to each combination of Interface, Block Reliability, and Trial Performance. An additional ART-ANOVA model was also used to measure (4) subjective task load scores (NASA-TLX), using only two predictors and their interaction: Reliability and Interface. This model did not contain Trial Performance, because NASA-TLX scores were collected at the end of each block, and therefore we could only consider block-level factors within this model. Effect sizes were calculated for each main effect using partial eta squared. Pairwise comparisons for significant main effects were carried out using contrasts from the 'emmeans' package, with Bonferroni corrections applied to account for multiple comparisons (Lenth, 2020). Visualisations were created using the 'ggplot2' R package (Wickham, 2016).

4.3.5.2 Data Availability

An anonymised version of this dataset will be made available by DOI through the UK Data Service ReShare repository. The UK Data Service is funded by the Economic and Social Research Council (ESRC) who provided funding for this project.

4.4 Results

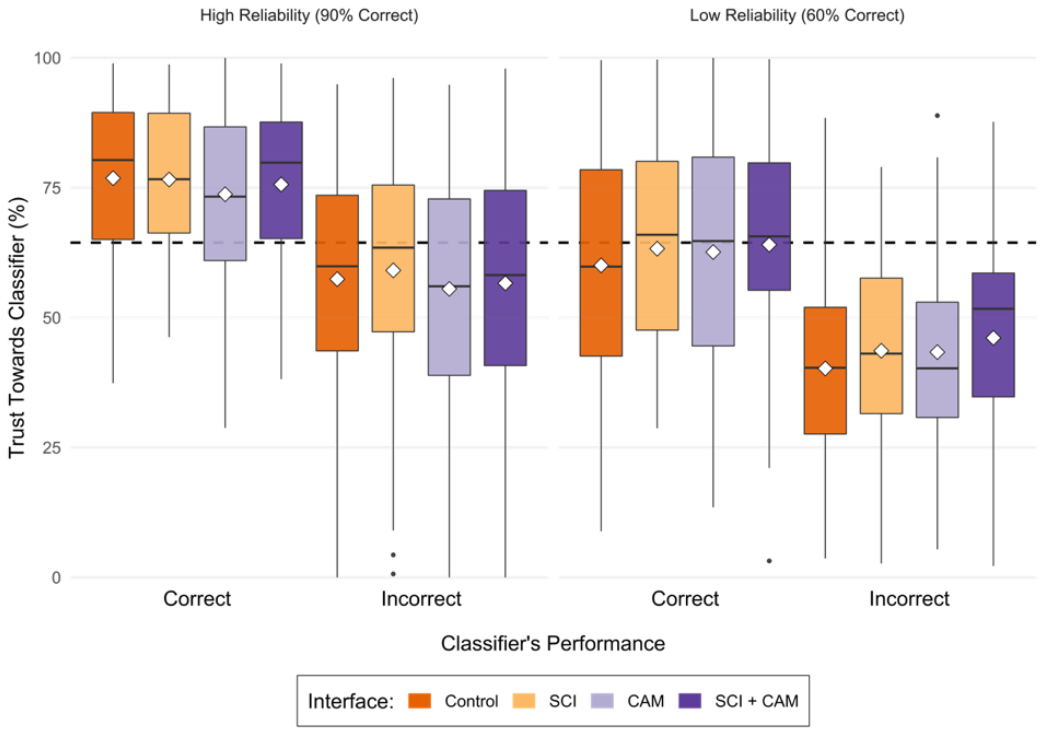
4.4.1 Influence of Interface on Trust

Participants' overall trust towards the classifier was highest when working with the SCI+CAM interface (M=65.38, SD=24.34), and lowest when working with the CAM interface (M=63.47, SD=25.31) (Table 4). The ART-ANOVA model for Trust revealed no

significant main effect of Interface F(3,144) = 1.04, p = 0.38, ηp2= 0.02. This suggests that there was no overall increase in trust associated with any of the interfaces. When examining trust within low reliability blocks, there appeared to be an increase in trust between the Control and SCI+CAM interfaces. When the classifier was correct in a low reliability block trust was lowest when working with the Control Interface (*M*=60.06, *SD*=25.66) and highest when working with the SCI+CAM interface (*M*=64.14, *SD*=23.47) (Table 5 and Figure 13). Likewise, when the classifier was incorrect in a low reliability block trust was also lowest when working with the Control Interface (*M*=40.13, *SD*=24.11) and highest when working with the SCI+CAM (*M*=45.96, *SD*=23.76) (Table 5 and Figure 13). This suggested there was some support for an interaction between Interface and Block Reliability, but this was not significant F(3,144) = 2.57, p = 0.06, ηp2= 0.05. Overall, this does not provide support for (H1a): Trust towards the classifier will be increased when working with the experimental interfaces. Likewise, this does not provide outright support for (H1b): Trust will be highest when working with the SCI+CAM interface, as it provides the most detailed information.

Figure 13

Trust Scores For Each Trial Performance Type At Each Reliability Level



Note. Dashed line represents grand mean for trust towards the classifier (%). White diamonds represent individual mean values for each condition.

4.4.1.1 Post-Hoc Power Analysis for Trust Model

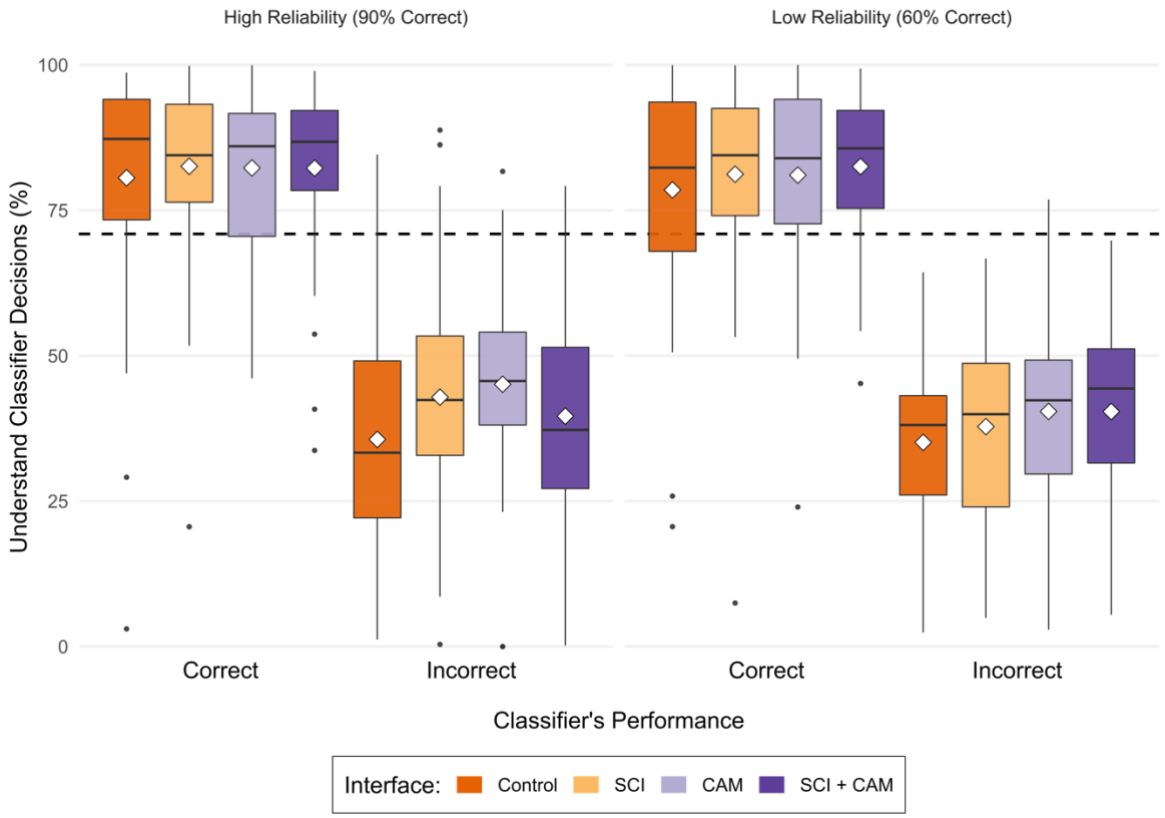
Power analysis for this study appeared similar to the power analysis conducted in Chapter 3, due to the similarity in the design of these experiments. Post-Hoc power analysis was conducted examining the influence of each of the 4 interface conditions on Trust. Our sample size for this experiment was 49 participants, with 4 observations (averaged values for each combination of trial performance and reliability) for each of the 4 interface conditions. The ART-ANOVA model for Trust Scores had a $\eta p^2=0.03$, suggesting we only had a 1-β power of 0.832. To achieve a 1-β power of 0.95 based on a within subjects design of 4 groups with 4 observations, we may have needed a sample of around 108 participants for full power in this study.

4.4.2 Influence of Interface on Understanding

Participants reported greatest understanding of the classifier's decisions when working with the SCI+CAM interface (*M*=71.94, *SD*=29.51) (Table 4). In contrast, participants typical understanding was lowest when working with the Control interface (*M*=68.72, *SD*=31.47) (Table 4). The ART-ANOVA model for Understanding revealed a small main effect of Interface $F(3,144) = 4.00$, $p < 0.05$, $\eta p2= 0.08$. Pairwise comparisons revealed that the only significant difference in understanding was between the Control and CAM interfaces ($p < 0.01$). There was also a significant interaction between Interface and Trial Performance $F(3,144) = 5.73$, $p < 0.001$, $\eta p2= 0.11$. This suggests that the extra information provided by the experimental interfaces may improve participants' understanding of the classifier's decisions, particularly interfaces featuring the CAM cue (Figure 14 and Table 6). This partially supports (H2a): Participants' understanding of the classifier will be increased when working with the experimental interfaces. However, this does not support (H2b): Participants' understanding will be highest when working with the SCI+CAM interface, which provided the most detailed information.

Figure 14

Understanding Scores For Each Trial Performance Type At Each Reliability Level



Note. Dashed line represents grand mean for understanding classifier's decision making (%). White diamonds represent individual mean values for each condition.

4.4.3 Influence of System Reliability

4.4.3.1 Trust

Trust towards the classifier was influenced by the classifier's performance in individual trials (Trial Performance), and the classifier's reliability throughout the block (Block Reliability) (Tables 5 & 7, Figure 15). The ART-ANOVA model for Trust revealed a significant interaction between Block Reliability and Trial Performance $F(1,48) = 10.68$, $p < 0.001$, $\eta p2 = 0.18$. There was also a main effect of Block Reliability $F(1,48) = 71.37$, $p < 0.001$, $\eta p2 = 0.60$, and a main effect of Trial Performance $F(1,48) = 41.86$, $p < 0.001$, $\eta p2 = 0.47$. This suggests that participants based their trust towards the classifier on both its performance within individual trials and its cumulative performance within the block. For example, in trials where the classifier was correct, trust towards the classifier was lower within low reliability blocks ($M = 62.54$, $SD = 24.36$), compared to correct trials within a high reliability block ($M = 75.69$, $SD = 18.55$) (Table 7). Conversely, trust towards the classifier was higher if an incorrect trial took place within a high reliability block ($M = 57.16$, $SD = 25.27$), compared to incorrect trials within a low reliability block ($M = 43.22$, $SD = 23.84$) (Table 7). Thus, evaluations of the classifier were informed by both the

performance within individual trials, and by the collective reliability across the block. This supports (H3): Trust towards the classifier will be highest in high reliability conditions.

Figure 15

Average Trust Scores For Each Image Used In The Experiment

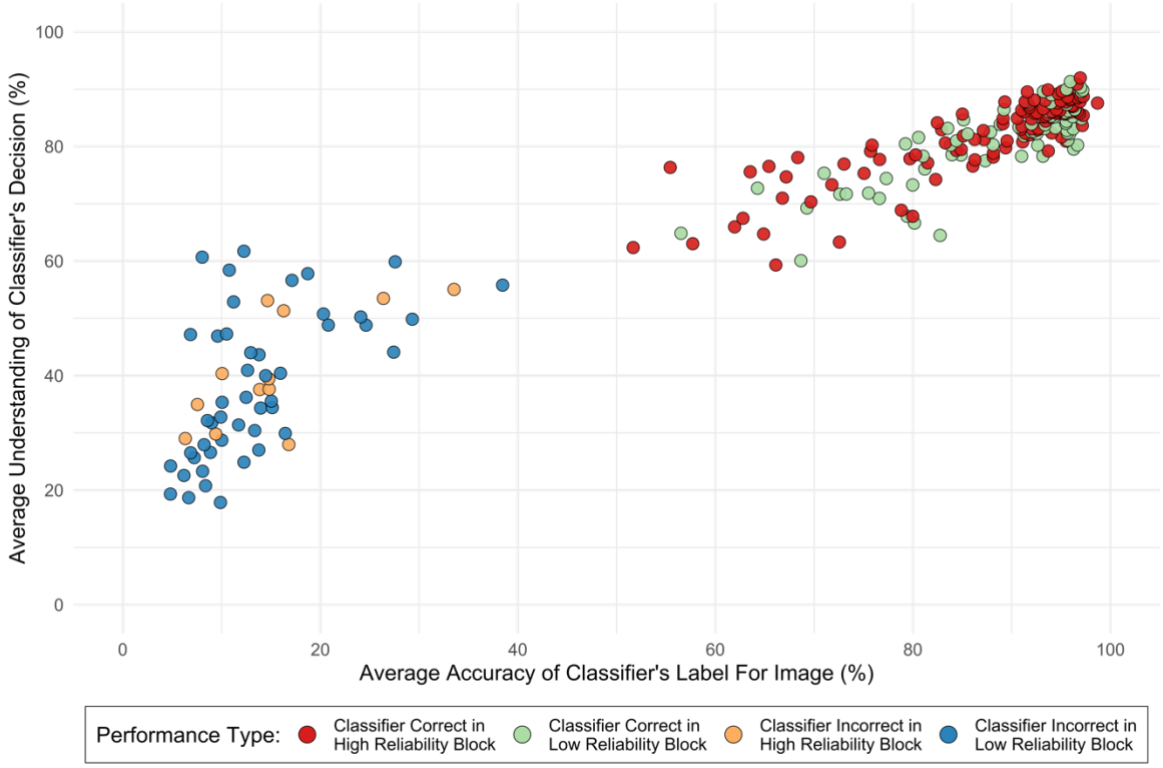
Note. Stimuli arranged by participants' average accuracy rating of classifier's label for the image.

4.4.3.2 Understanding

Participants' understanding of the classifier's decisions did not appear to vary greatly across reliability levels, instead their understanding was primarily based on the classifier's performance (Tables 6 and 8, Figure 16). Overall, when the classifier was correct, participants typically understood the classifier's decision more than when the classifier was incorrect (Table 8). The ART-ANOVA model for Understanding revealed a main effect of Trial Performance $F(1,48) = 275.37$, $p < 0.001$, $\eta p2 = 0.85$, while there was also a small but significant main effect of Block Reliability $F(1,48) = 4.63$, $p < 0.05$, $\eta p2 = 0.09$. There were no interactions between Reliability and Trial Performance. Participants' understanding of the classifier's decisions in correct trials was similar in both high reliability ($M = 81.93$, $SD = 21.98$), and low reliability conditions ($M = 80.88$, $SD = 21.93$) (Table 8). Likewise, there was limited difference in understanding scores for incorrect trials between high reliability ($M = 40.80$, $SD = 27.82$) and low reliability conditions ($M = 38.40$, $SD = 28.85$) (Table 8). This suggests that while participants understanding may

have been influenced by the classifier's reliability within the block, it was mostly based their understanding on the classifier's current performance within individual trials.

Figure 16
Average Understanding Scores For Each Image Used In The Experiment



Note. Stimuli arranged by participants' average accuracy rating of classifier's label for the image.

4.4.4 Participants' Task Load

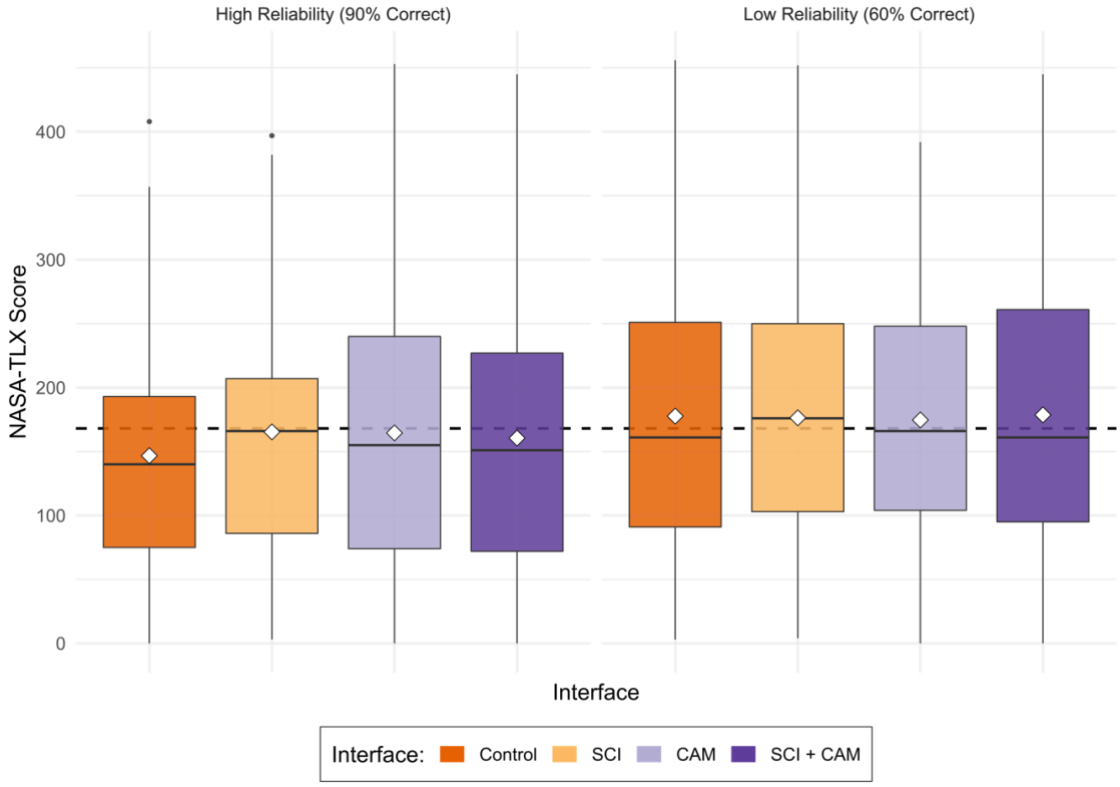
4.4.4.1 NASA-TLX

The classifier's reliability appeared to have the largest influence on participants' subjective task load scores, which were lower in conditions with high reliability, and higher in conditions with low reliability (Table 9, Figure 17). Across both high and low reliability conditions, task load was typically lowest when working with the Control interface and highest when working with the experimental interfaces (Table 4, Figure 17). The two-way ART-ANOVA model for NASA-TLX scores revealed a significant main effect for Block Reliability $F(1,48) = 15.29$, $p < 0.001$, $\eta p2 = 0.24$. However, there was no significant main effect for Interface $F(3,144) = 1.50$, $p = 0.22$, $\eta p2 = 0.03$. There was also no significant interaction between Block Reliability and Interface. This suggests that participants' subjective workload was increased in low reliability conditions, given they had to make more corrections to the classifier's labels. This provides support for (H4a): Participants' subjective task load scores will increase during low reliability conditions.

However, the addition of SCI, CAM or SCI+CAM cues did not appear to increase subjective workload, which does not provide support for (H4b): Participants' task load will be significantly higher when working with the SCI+CAM interface.

Figure 17

Subjective Task Load Scores For Each Interface At Each Reliability Level



Note. Dashed line represents grand mean for subjective task load scores (NASA-TLX). White diamonds represent individual mean values for each condition.

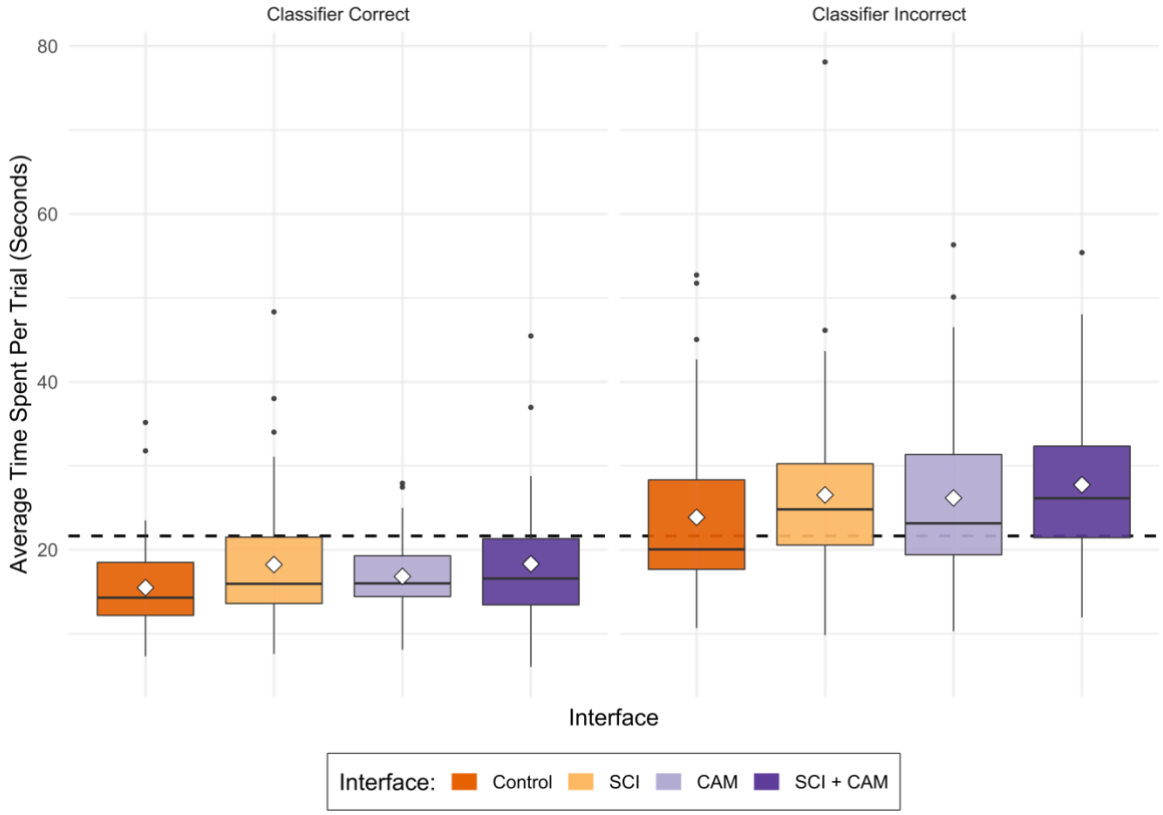
4.4.4.2 Trial Time

Participants typically spent the most time per trial (seconds) when working with the SCI+CAM interface (*M*=20.66, *SD*=22.02), and the least time per trial when working with the Control interface (*M*=17.58, *SD*=14.42) (Table 4, Figure 18). Time spent per trial was further influenced by the classifier's performance. Participants tended to spend more time in trials where the classifier's performance was incorrect (Table 10, Figure 18), which is unsurprising given participants had to correct the classifier's label in these trials. The ART-ANOVA model for Trial Time revealed significant main effects for Trial Performance $F(1,48) = 439.47$, $p < 0.001$, $\eta p2 = 0.90$ and Interface $F(3,144) = 10.46$, $p < 0.001$, $\eta p2 = 0.18$. There was no main effect for Block Reliability $F(1,48) = 0.37$, $p = 0.54$, $\eta p2 = 0.01$. There was however a 2-way interaction between Trial Performance and Interface $F(3,144) = 4.43$, $p < 0.05$, $\eta p2 = 0.08$. There were no other significant interactions within the Trial Time model. Pairwise comparisons revealed significant differences in time spent per trial was between the Control and SCI+CAM interfaces ($p<0.0001$), the Control and SCI

interfaces ($p<0.001$), and the Control and CAM interfaces ($p<0.01$). This suggests that participants tended to spend more time per trial when working with the classifier interfaces which provided the most information, and suggests that this additional information was not necessarily ignored by the participants. This provides some support for (H4b): Participants' task load will be significantly higher when working with the SCI+CAM interface.

Figure 18

Average Time Spent Per Trial With Each Interface For Correct and Incorrect Trials



Note. Dashed line represents grand mean for time spent per trial (seconds). White diamonds represent individual mean values for each condition.

4.5 Discussion

The primary purpose of this study was to explore whether trust towards an image classifier could be improved through displays of Gradient-weighted Class Activation Mapping (CAM) and Classifier's System Confidence Information (SCI). We also investigated whether these cues improved participants' understanding when working with the classifier, as a way of illustrating the decision-making of the system. We also explored the role that system performance, across blocks (reliability) and within individual trials (trial performance) played in shaping participants' understanding of system decision making and trust towards the classifier. Lastly, we also explored how the system's reliability and our cues of system decision making influenced users' task load.

4.5.1 Improving Trust Towards an AICS

Previous research suggests that increasing the transparency of system decision-making can improve users' trust towards autonomous systems (Chen et al., 2014; Mercado et al., 2016; Selkowitz et al., 2017; Tomsett et al., 2020). We provided participants with cues of SCI and CAM in a bid to increase the transparency of the AICS, yet there was no overall increase in participants' trust towards the classifier. This suggests that system reliability may have a much more profound impact on users' trust than the transparency of the system. This would mirror the findings of similar research, including our own (Chapter 3/ Ingram et al., 2021). Autonomous system accuracy was reportedly more influential on users' trust than the explanations provided by the system, when participants worked with an autonomous text classifier (Papenmeier et al., 2019). Likewise, when estimating their trust towards an autonomous squad member, the reliability of the system was reportedly more influential than the system transparency for users' evaluations (Wright et al., 2019).

We also expected that the SCI and CAM interface could improve participants' trust in low reliability conditions, as it provided the most detailed interpretation of system decision-making, which may be useful when there are more errors. In low reliability blocks, trust did appear to increase when users were provided the SCI and CAM cues together, relative to the control interface, yet this was not a significant increase (Table 5). However, we believe it is possible that with a more appropriate sample size there may be enough power to support this trend within the data. By contrast, we did not see this trend under high reliability conditions, where the Control interface tended to have the most trust (Table 5). Again, this may make sense when considering that in high reliability blocks the classifier was correct in most of the trials (90%), and so the additional information may have been seen as superfluous given the system was almost always correct. Thus, while SCI and CAM cues may not have increased overall trust, it remains possible that they may still be beneficial for improving trust towards an AICS if presented together during low reliability conditions.

4.5.2 System Reliability

As in our pervious experiment, participants' trust towards the classifier was mostly influenced by the performance of the classifier (Ingram et al., 2021). In the current study, the significant interaction between Trial Performance and Block Reliability suggests that participants modulated their trust in response to the classifier's performance within individual trials, yet also factored in its overall reliability within the block. This is illustrated most clearly within Figure 15, in which we looked at average trust scores for

each trial, and can see 4 clusters representing the 4 main combinations of Trial Performance and Block Reliability: (1) *Correct Trials in High Reliability Blocks*; (2) *Incorrect Trials in High Reliability Blocks*; (3) *Correct Trials in Low Reliability Blocks*; and (4) *Incorrect Trials in Low Reliability Blocks*. These results suggest that participants calibrated their trust differently for an incorrect performance within a high reliability block than they did within a low reliability block, and vice versa for correct trials. Ultimately, it appears the reliability within the block has a bigger influence on participants' trust towards the classifier, and suggests that participants based their trust more on the classifier's collective performance across the trials within blocks, rather than on individual trials. These results are not particularly surprising, as they replicate the findings of previous studies in which system reliability significantly influenced trust towards automation (Chavaillaz et al., 2016; Sauer & Chavaillaz, 2017; Hussein et al., 2019; Papenmeier et al., 2019; Wright et al., 2019). More importantly however, they develop upon the findings of our previous experiment, and reinforce the importance of system reliability when measuring trust towards autonomous systems.

In Ingram and colleagues (2021) (Chapter 3) none of our cues of SCI improved trust towards the AICS, despite them increasing the transparency of system decision-making. We speculated that the low overall reliability of the classifier within our previous experiment (50%), due to our experimental design, may have limited the potential benefits of SCI for improving trust. Within the current study, over 4 blocks of high reliability (90%) and 4 blocks of low reliability (60%) the classifier had a higher overall reliability (75%), Ironically, it is possible that within the current study our Interface cues may not have increased trust due to a ceiling effect within the high reliability (90%) condition. During high reliability blocks, participants trust for all 4 interfaces was typically near the upper limit of the scale we used to measure trust. It is therefore possible that during high reliability conditions, trust towards the experimental AICS interfaces could not be elevated any further due to the powerful effect of system reliability. Alternatively, it is equally possible that during high reliability conditions, the system was so reliable that the additional information provided by SCI and CAM cues became irrelevant, and therefore did not improve trust in a meaningful way. Nonetheless, the benefits of cues such as SCI and CAM may go beyond trust towards AICS, and these cues may also help to improve users' understanding of AICS decision-making.

4.5.3 Improving Users Understanding of AICS Decisions

While there is a great deal of research that is interested in improving trust towards autonomous systems, there is a parallel strand of research that is also interested in

improving the explainability of autonomous systems (Rudin, 2019; Gilpin et al., 2019). A common criticism of autonomous systems, particularly those based on neural network technology, is that these machines are uninterpretable 'black box' systems, in which the decision-making of the system is difficult, if not impossible, for the user to understand (Abdul et al., 2018; Ribera & Lapedriza, 2019). While we were primarily interested in seeing whether SCI and CAM cues can improve trust towards an AICS, we also explored whether they improved participants' understanding of classifier decisions.

Participants' understanding of the classifier was significantly improved when they worked with the CAM interface, yet there was no significant increase in understanding for the SCI or SCI+CAM interfaces. This would suggest that participants felt they could best interpret the decision-making of the AICS when working with the CAM cue. In theory this makes sense, as CAM appears to be useful for highlighting the areas and features of the image which influenced the classifier's decision. This supports Chen and colleagues (2019) suggestion, that cues such as CAM can visually represent the decision-making of AICS in a way that human users can easily understand. However, as Rudin (2019) suggests, there remains a disparity between CAM showing which parts of an image were influential for an AICS decision, and the user actually *knowing* what the AICS did when interpreting the features highlighted by CAM. A user could see the CAM cue highlighting features within an image that were relevant to the classification, yet still not grasp *why* these features are relevant to the AICS. Nevertheless, for the images that we used in our stimuli, our CAM cue appeared to be particularly useful at highlighting whether or not the AICS had detected the object of interest within the image. It therefore remains possible that participants' understanding was only enhanced by knowing whether or not the AICS had properly detected the object within the images. This is something which could also be illustrated by bounding box cues instead of CAM, which may take less effort to interpret, whilst still highlighting areas of interest. Future studies may benefit from examining more simplified ways of illustrating AICS activity when processing images, particularly if these cues can improve trust and/or understanding without increasing the user's workload.

4.5.4 Workload

When implemented within real-world settings, cues of SCI and CAM may improve AICS transparency. However, these cues will only be beneficial if they do not overwhelm the operator, and may not be used if they are detrimental to the user's productivity. If the interface is well designed, increasing the transparency of an autonomous system's interface may not automatically increase a user's workload (Mercado et al., 2016; Selkowitz et al., 2017). When examining subjective task load through NASA-TLX questionnaires, our

participants reported no significant increase in perceived workload when working with the experimental interfaces. We had expected to see an increase in subjective workload when they worked with the SCI+CAM interface, but this was not the case. Based on the current trends within the data (Table 4 and 9) we believe there may have been a significant difference in subjective task load if we increased our sample size. Interestingly however, we did see an increase in the average time spent per trial when participants worked with all 3 of the experimental interfaces. The biggest increases in average time spent per trial were seen in the SCI and SCI+CAM interfaces, which makes sense as these are the most 'information-heavy' versions of the classifier's interface. This also suggests that participants did not ignore the extra information presented in the experimental interfaces. Moreover, this also aligns with their ratings for their preferred interface (Table 4). When asked which interface they preferred the majority picked the SCI+CAM interface (44.9%), with the SCI-only interface the next most popular (30.6%) (Table 4). Thus, while participants spent more time per trial with the experimental interfaces, they also overwhelmingly preferred working with them, suggesting they found them beneficial.

Ultimately, users' workload appeared to be most significantly impacted by the classifier's performance, both within individual trials and through changes in its reliability between blocks. These findings replicate those of similar studies, in which imperfect automation can influence the workload of the operator using the system, by leaving the user with a larger share of the task to complete themselves (Sauer et al., 2016; de Visser & Parasuraman, 2011). The largest increases we saw in subjective workload were between the high and low reliability conditions. Likewise, participants' average time spent per trial was most significantly increased in trials where the classifier's label was incorrect. Neither finding should be seen as surprising, participants were expected to take more time in trials where they had to correct the classifier's errors, and would likely report a higher subjective workload in low reliability blocks where there were simply more errors to correct. Thus, while cues such as SCI and CAM may improve AICS transparency, their effect on workload appears minimal in comparison to the overall performance of the system.

4.5.5 Future Directions and Limitations

Improving system transparency appeared to be beneficial for users working with an AICS, in particular the CAM cue which appeared to improve understanding. This suggests that CAM may be particularly well suited to further research investigating explainability and trust towards AICS. Future research may benefit from exploring the utility of these cues as an optional resource which can be manually accessed by the user. Within the blocks where they were made available, CAM and SCI were automatically presented to the

user as part of each trial. It may be useful to explore participants use of SCI and/or CAM when they are an optional resource that can be manually accessed within each trial, as opposed to something that is automatically presented to the user. This could give a clearer indication of whether or not users actually want to use these cues when working with an AICS. Making these cues optional could also limit the impact they have upon the user's workload, by presenting them only when requested.

Undoubtedly our interpretation of these results was impacted by the limited sample of participants that we were able to collect due to the Covid-19 pandemic. For the purposes of this PhD, we analysed and reported the results as if we had completed our data collection, but would have preferred to have a sample size similar to the sample used in Chapter 3 (n=74).

4.6 Conclusion

Our study sought to build upon the findings from our previous experiment, by exploring whether we could improve users' trust when working with an Autonomous Image Classifier System (AICS). For the most part, trust was primarily influenced by the general performance of the classifier: whether the classifier's label correctly or incorrectly described the image. Changes in participants' trust appeared to be most significantly linked to changes in the reliability of the system across experimental blocks: increasing in high reliability blocks, and lowering in low reliability blocks. We also examined the usefulness of both Gradient-weighted Class Activation Mapping (CAM) and Classifier Confidence Information (SCI), as a way of improving trust towards the classifier. We did not find an increase in trust towards the AICS when participants worked with the interfaces featuring SCI and/or CAM. We did however find that participants' understanding of the classifier's decision-making appeared to be significantly improved when working with the interface featuring the CAM cue. While ratings of subjective workload were not increased when working with the experimental interfaces, there was an increase in the average time spent per trial for all 3 experimental interfaces, relative to the control. Future research may be able to build upon these findings, in order to further explore the utility of these cues of system decision-making.

4.7 Data Tables

Table 4

Descriptive statistics for Trust, Understanding, Compliance, Identifiability, Time, TLX, Aesthetics and Overall Preference with each Interface.

	Trust the Classifier (%)		Understand Classifier (%)		Label Compliance (%)		Identifiability of Images (%)		Time Per Trials (Seconds)		Task Load (NASA-TLX)		Aesthetic Rating (1-7)		Favourite Interface
Interface	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*N (%)*
Control	63.59	26.22	68.72	31.47	73.28	44.26	82.29	23.23	17.58	14.42	162.20	102.95	4.54	1.58	8 (16.3%)
SCI	65.26	24.25	71.36	29.96	72.51	44.66	82.00	24.16	20.29	21.86	170.86	100.69	4.70	1.42	15 (30.6%)
CAM	63.47	25.31	71.77	29.27	73.20	44.30	82.13	23.71	19.15	15.36	169.61	102.79	4.79	1.40	4 (8.2%)
SCI+CAM	65.38	24.34	71.94	29.51	72.86	44.47	81.80	24.24	20.66	22.02	169.58	109.04	4.83	1.34	22 (44.9%)

Note. *M* and *SD* represent mean and standard deviation, respectively.

Table 5

Means and standard deviations for Trust (%) as a function of a 4(Interface) X 2(Reliability) X 2(Trial Performance) design

	High Reliability						Low Reliability					
	Correct		Incorrect		Overall		Correct		Incorrect		Overall	
Interface	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Control	76.84	17.98	57.41	26.65	74.89	19.89	60.06	25.66	40.13	24.11	52.12	26.88
SCI	76.57	17.72	59.09	23.98	74.82	19.16	63.32	23.07	43.57	23.03	55.47	24.99
CAM	73.74	19.57	55.53	25.64	71.91	20.97	62.66	25.01	43.22	24.15	54.91	26.44
SCI+CAM	75.63	18.75	56.62	24.86	73.73	20.26	64.14	23.47	45.96	23.76	56.90	25.20

Note. M and *SD* represent mean and standard deviation, respectively.

Table 6

Means and standard deviations for Understanding (%) as a function of a 4(Interface) X 2(Reliability) X 2(Trial Performance) design

	High Reliability				Low Reliability			
	Correct Trials		Incorrect Trials		Correct Trials		Incorrect Trials	
Interface	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Control	80.59	24.31	35.61	29.03	78.48	22.99	35.18	27.80
SCI	82.58	21.22	42.89	26.64	81.25	22.23	37.72	28.56
CAM	82.29	20.84	45.10	26.63	81.16	21.93	40.32	29.56
SCI+CAM	82.26	21.36	39.62	28.26	82.64	20.33	40.37	29.21

Note. M and *SD* represent mean and standard deviation, respectively.

Table 7

Means and standard deviations for Trust Score as a function of a 2(Trial Performance) X 2(Reliability Level) design

	Reliability Level			
	High Reliability		Low Reliability	
Performance in Trial	*M*	*SD*	*M*	*SD*
Correct	75.69	18.55	62.54	24.36
Incorrect	57.16	25.27	43.22	23.84

Note. M and *SD* represent mean and standard deviation, respectively.

Table 8

Means and standard deviations for Understanding Score as a function of a 2(Trial Performance) X 2(Reliability Level) design

	Reliability Level			
	High Reliability		Low Reliability	
Performance in Trial	*M*	*SD*	*M*	*SD*
Correct	81.93	21.98	80.88	21.93
Incorrect	40.80	27.82	38.40	28.85

Note. M and *SD* represent mean and standard deviation, respectively.

Table 9

Means and standard deviations for Task Load (NASA-TLX Scores) as a function of a 4(Interface) X 2(Reliability) design

	Reliability Level			
	High Reliability		Low Reliability	
Interface	*M*	*SD*	*M*	*SD*
Control	146.71	95.82	177.69	108.38
SCI	165.35	98.24	176.37	103.80
CAM	164.53	109.39	174.69	96.60
SCI+CAM	160.51	103.20	178.65	114.93

Note. M and *SD* represent mean and standard deviation, respectively.

Table 10

Means and standard deviations for Trial Time (Seconds) as a function of a 4(Interface) X 2(Trial Performance) design.

	Trial Performance			
	Correct Trials		Incorrect Trials	
Interface Name	*M*	*SD*	*M*	*SD*
Control	15.48	12.89	23.93	16.78
SCI	18.24	22.69	26.52	17.72
CAM	16.81	14.20	26.21	16.54
SCI+CAM	18.33	22.74	27.74	17.94

Note. M and *SD* represent mean and standard deviation, respectively.

Table 11

Descriptive statistics for responses to Debriefing Questionnaire questions (1-6, and 8)

	Response (1-7)	
	M	*SD*
How helpful did you think the classifier was? <Not at all / A Great Help>	4.9	1.0
How predictable was the classifier's behaviour? <Predictable / Unpredictable>	3.5	1.3
How specific did you think the classifier's labels were? <Too Specific / Too General>	3.1	1.3
If you had to describe it to someone, how you would characterise the classifier? <Teammate / Tool>	5.6	1.4
If you had to classify another set of images, would you want to work with the classifier again? <With Classifier / Alone>	2.7	1.4
If you had to classify another set of images, which type of collaborator would you prefer? <Computer / Human>	4.5	1.9
Overall, how well would you say you typically understand technology? <Expert / Novice>	3.4	1.5

Note. *M* and *SD* represent mean and standard deviation, respectively.

Chapter 5: General Discussion

5.1 Summary of Study Findings

5.1.1 Summary: Qualitative Exploration of Trust Towards Technology.

Our first study sought to explore trust towards technology from a qualitative perspective. A recurrent theme that became apparent during my initial literature review, was that the majority of the existing research was being carried out using quantitative tools, from researchers who were predominantly from computing science and engineering backgrounds. As such, I felt that there was a gap within the literature exploring how users characterize trust towards technology. Through focus group interviews, we asked participants to explain their thoughts on popular technologies, such as Social Media Services and online retailers. Even though this research did not involve direct interaction with the autonomous technologies frequently seen in other literature, these findings still appeared to be relevant. Participants expected their information to be secure when using these services, and a violation of this expectation, through accidental leaks or malicious agents, was interpreted by many as being fatal for trust towards the service. Likewise, many participants (but not all) were interested in how their information was used by these services, which could be interpreted as a need for greater transparency within these services. These findings were also potentially linked to the Cambridge Analytica Scandal which was exposed in the months prior to this research (Cadwalladr & Graham-Harrison, 2018), which likely prompted many participants to revaluate their relationship with these technologies.

5.1.2 Summary: Calibrating Trust Towards an Autonomous Image Classifier

Our second study sought to investigate trust within an experimental context. We explored how human users calibrate their trust towards an Autonomous Image Classifier System (AICS) when completing an image identification task. By doing so we provided, to our knowledge, the first insight into how human users calibrate their trust specifically towards an AICS. We found that trust was primarily based upon the system's performance, if the AICS was correct trust was higher, and when incorrect trust was lower. We also demonstrated how environmental factors further influenced trust calibration. When processing low quality images (blurred and cropped contents) participants were more cautious when estimating their trust, and tended to report values closer the midpoint of our trust scale. By comparison, participants were much more extreme with their trust

evaluations when processing higher quality (clearer) images. This illustrated how human users considered the difficulty of the task facing the AICS when estimating their trust.

This design choice was heavily influenced by Merritt and colleagues (2013) who examined trust towards a weapons detection system when processing images of luggage which were either empty or full/cluttered. This research stood out to me because it was one of the few studies which fully explored a scenario that the human-machine team may face within real-world settings. In such settings, performance of the autonomous system becomes much harder to appraise, as the success of failure of the system is not so readily apparent, and this may influence how the operator relies upon a system. I believe that introducing these scenarios within laboratory experiments may improve the ecological validity of the research by limiting the capacity for the human user to carry out the task themselves. I was inspired to try and follow this example by manipulating the quality of the images processed by the AICS within this experiment. In these trials the performance of the classifier would also be harder to evaluate, since the participant cannot as easily identify the contents of the image themselves. This could have also created an opportunity where some participants became more reliant on the system, if the labels it provided plausibly matched the blurred contents of the image. Ultimately, I believe the decision to have unclear images provided greater insight into how participants placed trust in the system, and also prevented participants from completing the entire task themselves without considering the advice of the classifier.

The other main component of this research was exploring whether trust towards the AICS could be improved by making the system more transparent through different displays of System Confidence Information (SCI). A great deal of existing research demonstrates that trust towards autonomous systems is typically improved by making the system's decision making more transparent (Chen et al., 2014; Desai et al., 2013; Mercado et al., 2016; Selkowitz et al., 2017; Tomsett et al., 2020; Zhang et al., 2020). Somewhat surprisingly, we did not see a significant increase in participants trust when provided with cues of SCI, suggesting this may not be an appropriate way to improve trust towards an AICS. This was disappointing, as we thought the provision of SCI would help to better illustrate the decision making of the AICS, particularly when the AICS was processing low quality images, where it may be harder to evaluate the system's performance. Instead, there was no observable increase in participants' trust towards the classifier, yet when asked for their favourite interface, participants overwhelmingly preferred the interfaces with SCI. This suggests they found the extra information beneficial on some level, and warranted further exploration.

5.1.3 Summary: Promoting Understanding and Trust Towards an Autonomous Image Classifier

Our third study directly followed on from the findings in Chapter 3, by examining new ways to improve trust towards an AICS. In Chapter 3, whilst participants preferred interfaces that provided SCI, the cues we provided did not significantly improve trust towards the AICS. We speculated that there were a variety of potential reasons for this lack of impact on participants' trust and sought to test these theories in our final study. We first modified our experimental paradigm so that the classifier would have a higher reliability throughout this experiment, given the importance of system performance reliability (Chavaillaz et al., 2016; Wright et al., 2019; Yu et al., 2019). We also introduced an entirely new method for improving system transparency: Gradient-weighted Class Activation Mapping (Grad-CAM/CAM) (Selvaraju et al., 2017; Yang et al., 2019). Through CAM, we were able to visually represent the activity of the AICS when processing images, thereby highlighting the regions and features that were important for each classification. By doing so, we could then compare the utility of both SCI and CAM as way of improving trust through increased system transparency. We also recognised that it was equally possible that these cues may not actually increase trust. Therefore, as additional factor of interest, we also looked at the whether these cues improved users' understanding of the AICS decision making. This was because we speculated that it was also possible that participants may prefer these cues simply because they help them to better understand the machine's decisions, without making it more trustworthy.

While our data collection was disrupted by the COVID-19 pandemic, we were still able to find some interesting insights within our preliminary data. Participants overwhelmingly preferred the most complicated version of the interface (SCI+CAM), and while participants typically spent the most time working with this interface, they didn't report an increase in their subjective workload. This would suggest these cues may be useful to users without over encumbering them with information. During low reliability blocks trust towards the classifier was also highest when working with the SCI+CAM interface, yet this was not a significant increase relative to the control interface. A larger sample size may clarify whether there was a true increase in trust with these cues. Additionally, we found that participants' understanding of the classifier's decision making was increased by the presence of the CAM cue. These findings show support for the utility of the CAM cue improving users' confidence towards an AICS.

5.2 Contributions to the Field

5.2.1 Measuring Trust Towards Automation

The central aim of the research within this thesis was to explore trust towards technology and automation. A criticism of this research, and indeed the wider trust-in-automation literature could be that a user's trust towards an autonomous system may simply be their estimation of reliability, or their confidence in the system's performance. Much of the existing literature places system reliability and performance as central factors in trust towards automation. This should not be surprising, given that autonomous systems are typically designed to complete a specific set of tasks, and any deviation from this represents a violation of their fundamental purpose. Across all three of my studies, users' trust was also primarily informed by the performance of the technology: a well-functioning machine merits trust, whilst an underperforming machine invokes distrust. This is consistent with a considerable amount of existing literature, in which trust towards autonomous systems is closely linked to system performance and reliability (Chavaillaz et al., 2016; Desai et al., 2013 Sauer & Chavaillaz, 2017; Hussein et al., 2019; Papenmeier et al., 2019; Parasuraman & Riley; 1997; de Visser et al., 2018; Wright et al., 2019; Yu et al., 2019).

In Chapter 2, during focus group discussions, participants touched on a variety of factors that may shape and inform their trust towards digital online technologies. Participants appraisals of these services were influenced by the security of their information on these services, as well as their perception and understanding of how their information is used by these services. Both factors can be considered as important to the performance of these services. If users' information is not secure, or at least does not appear to be secure, these technologies are not completing performing a fundamental task expected of their services. Similar research suggests if people do not feel that they have control over the information they share with these services, they may be less likely to share more information (Benson et al., 2015). This in itself, may affect the performances of these services, given that the entire digital ecosystem is reliant on users' willingness to have their data collected. Equally, as users are required to share their information with these digital services, if there is a lack of clarity in how this information is used, these services may need to convey their processes to users in a clearer manner. It is easy to see how this could negatively impact upon users trust when it is not clearly communicated. For example, Jung (2017) reports that users had greater concern about their privacy and their information

when they were exposed to targeted ads on social media. During our focus groups, participants interpretations of how their information was used by these services, and the extent to which their information was harvested, was interpreted in many different ways. The bigger picture was that a lot of participants spent time discussing this, and regardless of experience or background, believed that the communication of these practises was important for fostering trust. Moreover, some participants showed an awareness of mitigating factors that could also inform their trust towards these services. Some participants highlighted their lack of knowledge or interest in digital technologies as something that shaped their trust (or lack thereof) towards these services, while others noted the impact that other humans could have, particularly in the case of social media services. Therefore, the measurement of trust in this study provides a unique insight into how users may characterise their trust towards technologies, and provides perspectives from participants which appear to align with evidence from existing literature

In the experiments within Chapters 3 and 4, trust was measured using a variety of methods. The most straightforward method used participants' scores from the Propensity to Trust Machines Questionnaire (PTMQ), which used 6 questions exploring how likely they were to trust new technologies. This was collected once per participant, either at the start or at the end of the experiment. By doing so, this captured participants' trust towards technology as a static, trait-like measurement. In Hoff and Bashir's 2015 model, PTMQ scores could be considered as a form of Initial Learned Trust, which is based on users' previous interactions with technology. Similar to this, Lee and See (2004) also recognise the users' 'predisposition towards trust' as a component of the 'trust evolution' process, wherein users' trust towards a system is shaped by continuous interaction with the system. Secondly, as participants interacted with the classifier in both of our experiments, in each trial they were also asked to update their trust towards the classifier, to reflect how much they would want to continue working with the classifier. In each trial they also had to decide, ultimately, whether or not to reject the classifier's suggested label for each trial, which was termed as Compliance. Within Hoff and Bashir's (2015) these measures of trust and compliance could be considered as Dynamic Learned Trust, given they are based on continuous interaction with the classifier, and represent a more state-like measurement of trust. In Chapter 3, we found correlations between participants PTMQ scores and their trust towards the classifier, as well as between their PTMQ scores and their Compliance with the classifier. This suggests that our measurement of trust within each trial aligned with our measurement of trust using the PTMQ questionnaire. This also mirrors other findings within the literature where higher scores of Automation Bias and Propensity to Trust

Machines correlate with users placing more trust in autonomous technologies (Merritt et al., 2013; Goddard et al., 2014). Ultimately, we demonstrate that trust towards an autonomous image classifier can be measured on a trial-by-trial basis, and that this measurement can align with other measures of trust, such as the PTMQ.

5.2.2 Improving Trust Towards Automation

Within the context of Bonnie Muir's (1987) exploration of trust towards automation, the research in Chapters 3 and 4 has primarily focussed on her first suggestion: "Improving the user's ability to perceive a decision aid's trustworthiness". In Chapter 3, we used different cues of SCI as a way of exploring their potential for improving the transparency of an autonomous image classifier system. In Chapter 4, we also introduced Grad-CAM as an alternative way of improving system transparency. Both cues were intended to improve our participants capacity for interpreting the decision making of the classifier. The Situation awareness-based Agent Transparency (SAT) model (Chen et al., 2014), suggests that autonomous system transparency can be improved by providing information about system performance within the interface. Within the SAT model, providing more information should promote greater transparency, and by extension more trust from the user, in keeping with Muir's (1987) suggestion. However, in both studies our results were less than conclusive, trust was not directly improved by the presence of SCI or Grad-CAM cues, even though participants preferred working with the interfaces which provided more information. Previous literature employing SCI-type cues have found that they improve trust towards other types of automation (Desai et al., 2013; McGuirl and Sarter, 2006; Verame et al., 2016; Zhang et al., 2020). As previously speculated, the lack of improved trust in our experiments may have been linked to our experimental design, or could be due to SCI being less relevant when working with an AICS. Nonetheless, in science Null Results remain important, and may still contribute towards our understanding of trust towards automation. Ultimately, I believe the main benefit of this research was comparing different types of SCI together, and doing so across multiple types of system performance. Much of the existing literature has only looked at these factors separately, and this research has attempted to bring these factors together. While our results may suggest that neither SCI or CAM are appropriate for improving trust towards an AICS when used alone, additional factors may also need to be considered

Instead, we may need to look again to Muir's (1987) suggestions for other routes to evaluate these cues. For example, when "Improving the user's ability to perceive a decision

aid's trustworthiness", we used cues such as SCI and CAM. However, this suggestion could be also done through additional training for our participants/users. While we did have a limited training block at the start of our experiments to get participants accustomed to the classifier's interface, we may have benefited from a more detailed training session which could have explained to users how the AICS works, and go into detail on how cues such as SCI and Grad-CAM can explain the system's decision-making parameters. This may have also allowed for us to engage with Muir's second suggestion: "Modifying the user's criterion of trustworthiness", where we could have helped to define and illustrate the performance and reliability of the AICS to the participants before using the system. While this may have made participants more aware of how competent the AICS was, there is also the danger of biasing participants' opinions toward the system before they've even had a chance to interact with it. In doing so, participants could have based their trust scores primarily on the initial training we provided, rather than using evidence of directly interacting with the classifier in the experiments. If users are exposed to an autonomous system that does not commit any errors during training, this can increase users' Automation Bias, making them overestimate the capabilities of the system (Sauer et al., 2016). Similarly, if the reliability of the system is low during training, participants are less likely to trust the system when working with it (Chavaillaz & Sauer 2017). Thus, any attempt to train participants prior to interaction with the system should take serious consideration of how best to inform participants without misleading them about performance, and without overwriting their 'raw perceptions' of the system. Something like this could be tested using a between-subjects experimental design, allowing for comparison of trust scores of participants with limited versus detailed training.

Training could also be used for Muir's (1987) other suggestion: "Identifying and selectively recalibrating the user on the dimension(s) of trust which are poorly calibrated". Essentially, we may have been able to use participants' PTMQ or Automation Bias (Mosier et al, 2017) scores, to identify and provide training to participants with excessively low or excessively high predispositions towards trusting machines. These individuals may have benefited more from training framing the system's competence, thereby helping them to calibrate their trust more effectively. Such an approach could also be beneficial within real-world settings when introducing an automated system to a workplace. If a costly new autonomous system is introduced to a workplace, it would make sense to examine which potential users are more prone to under-trusting or over-trusting the system, and provide them with further training to recalibrate their perceptions. Lastly, while Muir (1987) also suggests "Enhancing the user's ability to allocate functions in a system", in our

experiments participants always had the ultimate say on the final label for each image. This means that participants were always in control of the final decision in each trial, giving them significant authority over the AICS. Theoretically, a way to engage with this suggestion may have been to allow participants to change the classification model, which may have then provided different classifications for certain images, and provided participants with a greater sense of control over the AICS. Such a feature would have been difficult to control for within our experimental conditions – but may be more feasible within an experiment designed to specifically test this.

5.3 Limitations

Our experiments with the AICS in Chapters 3 and 4 used the pretrained AlexNet and SqueezeNet models, around which I built the rest of the classifier's interface. By doing so, we used these models in a 'plug and play' capacity, in which we did not retrain or recalibrate the decision making of the classifier. As a result, some of the labels output by the classifier could be considered as linguistically strange. For example, it would likely classify an image of a white bear in a snowy environment as an 'Ice Bear' rather than a 'Polar Bear'. I tried to circumvent this by being selective with the stimuli I used in these experiments, by picking only images which were given more conventional labels by the classifier. We also asked participants to rate the accuracy of these labels as a way of cross-validating the appropriateness of the decisions made by AlexNet and SqueezeNet. In an ideal world, I would have liked to have taken my own set of images, and then trained the classifier to recognize the contents of them. I believe by doing so we would have had greater levels of control over the output of the classifier, and could be absolutely certain about whether the classifier's decision was 'truly correct'.

In the end, for the sake of convenience, I used stimuli taken from the Open Images Dataset(s) V4 and V5 (Kuznetsova et al., 2020), and used the attached labels within the dataset as the 'true' label for the image. The labels output by AlexNet and SqueezeNet were then compared to these 'true' labels, as a way to ascertain Correct vs Incorrect performances from the AICS. Even this strategy was not 100% perfect, as there were some images in the Open Images Dataset that I didn't feel I could use in good conscience. The most memorable instance of this was an image of a big cat which was defined in the Open Images Dataset as a 'Jaguar', but SqueezeNet defined it as a 'Leopard'. A cursory Google image search only added to the confusion, and I could foresee participants having similar issues in determining whether the classifier was 'truly correct' in its decision making. Ultimately images where I significantly doubted the label provided by the Open Images

Dataset were not used in our experiments. If we had been able to train the classifier using our own set of images, I believe I could have been absolutely certain about the appropriateness of the labels the classifier provided in our experiments.

Additionally, while we explored trust using quantitative and qualitative methods, most of this research relied upon subjective reporting from participants. If I had to do it all again, I think I would have tried to integrate a behavioural measurement of trust within some of our research too. The idea of implementing a tool such as Eye Tracking was always in the back of my mind, and I believe this would have been useful for our experiments with the AICS. It would have been interesting to explore whether users' fixations changed in response to fluctuations in the classifier's performance. Furthermore, we may have gained a better insight into the utility of the different interfaces we provided in Chapters 3 and 4, by seeing whether participants truly looked at the SCI and CAM cues. Ultimately, I was unable to use any behavioural methods such as Eye Tracking due to a lack of time, but I firmly believe it would be worth exploring in future research.

5.4 Future Directions & Closing Remarks

As suggested above, I would like to see more objective measurements of trust within our field. I believe that behavioural methods such as eye tracking could improve our understanding of trust towards technology. In a similar vein, when exploring the utility of cues such as SCI and CAM, I think it would be interesting to have these cues available as optional resources which are manually accessed by the participant. In our experiments these cues were always displayed automatically within each trial, but I think the true usefulness of these cues could be demonstrated when participants have to deliberately choose to see them. It is possible that the user may choose to use these cues more when a system has low reliability, and may in fact ignore these cues when a system has high reliability.

Our findings are primarily centred on trust towards an AICS, but some of these findings may be generalisable to research involving other types of automation. We found correlations between trust towards the AICS and participants PTMQ scores, in line with previous research (Merritt et al., 2013; Goddard et al., 2014). In Chapter 3, we also demonstrated the influence of task difficulty on participants trust towards the autonomous system, which appeared to have a dampening effect on extreme trust scores, regardless of whether or not the classifier was correct or incorrect. This also mirrors some of the existing literature, where users are capable of factoring the difficulty of the task into their evaluations of automation (Goddard et al., 2014; Lyell et al., 2018; Merritt et al., 2013; Schwark et al., 2010). SCI is a popular cue of system decision making amongst other

autonomous systems, yet we demonstrate that it may not be suitable for improving trust in low reliability settings. We also demonstrated that participants' perceived understanding of the AICS performance was based on performance within individual trials, and not on the reliability of the system within the block. This does not necessarily mean that participants actually understood what the classifier was doing when it made these decisions, rather, participants believed they understood what it was doing. This is important, as it suggests that trust and perceived understanding of an autonomous system are not entirely dependent on each other – a system may have low reliability and therefore the user may distrust it, but they may still feel they can understand why it made mistakes. This would be worth exploring using other autonomous systems to see if this process can be fleshed out further.

Regarding real-world applications of AICS, I think the CAM cue seems like a particularly useful way of representing AICS decision making. In Chapter 4 we seen an increase in participants' understanding of the classifier's decision making. With a bigger sample size, it may have also benefited trust when combined with SCI in low reliability conditions. Stepping away from my role as a researcher and into the role of a user, I can definitely see why the CAM cue could be particularly useful. During the countless hours I spent designing the experiments in Chapters 3 and 4, I became significantly acquainted with the quirks and mannerisms of the AlexNet and SqueezeNet models. Whilst I still found SCI beneficial for interpreting the decision making of the AICS, I felt that the CAM cue was particularly intuitive to use. I found it took less effort to interpret CAM, and thought in many cases it was eye-catching and intriguing to look at when working with the classifier. Above all though, I feel that the CAM cue gives the user a much clearer insight into whether or not the classifier has actually 'seen' the object(s) of interest. By contrast I feel that the SCI cue, whilst useful, does not convey the machine's interpretation to quite the same degree, and may still come across as more of a black-box system as a result. In my humble opinion, CAM is therefore the cue with biggest potential for making image classifiers more intuitive and trustworthy for their users. I will watch on with interest to see whether the next generation of researchers can prove this to be the case.

Appendices

Debriefing Questionnaire – Chapter 3

1.How helpful did you think the classifier was?

<Not at all / A Great Help>

2.How predictable was the classifier's behaviour?

<Predictable / Unpredictable>

3.How specific did you think the classifier's labels were?

<Too Specific / Too General>

4.If you had to describe it to someone, how you would characterise the classifier?

<Teammate / Tool>

5.If you had to classify another set of images, would you want to work with the classifier again?

<With Classifier / Alone>

6.If you had to classify another set of images, which type of collaborator would you prefer?

<Computer / Human>

7.If you had to quickly classify another set of 1000 images, which version of the interface would you prefer?

<Numerical, Iconography, Graphical, or Control Interface>

Debriefing Questionnaire – Chapter 4

1.How helpful did you think the classifier was?

<Not at all / A Great Help>

2.How predictable was the classifier's behaviour?

<Predictable / Unpredictable>

3.How specific did you think the classifier's labels were?

<Too Specific / Too General>

4.If you had to describe it to someone, how you would characterise the classifier? <Teammate / Tool>

5.If you had to classify another set of images, would you want to work with the classifier again? <With Classifier / Alone>

6.If you had to classify another set of images, which type of collaborator would you prefer? <Computer / Human>

7.If you had to quickly classify another set of 1000 images, which version of the interface would you prefer? <Control, SCI, CAM, or Max Interface>

8.Overall, how well would you say you typically understand technology? <Expert / Novice>

9.If you could only choose one cue, which one would you choose? <Heat Maps (CAM) / Confidence Info (SCI)>

References

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018, April). Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. *In Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1-18). https://doi.org/10.1145/3173574.3174156

Abraham, H., Lee, C., Brady, S., Fitzgerald, C., Mehler, B., Reimer, B., & Coughlin, J. F. (2016). Autonomous vehicles, trust, and driving alternatives: A survey of consumer preferences. Massachusetts Inst. Technol, AgeLab, Cambridge, 1, 16. https://bestride.com/wp-content/uploads/2016/05/MIT-NEMPA-White-Paper-2016-05-30-final.pdf

Adhikari, V. K., Guo, Y., Hao, F., Hilt, V., Zhang, Z. L., Varvello, M., & Steiner, M. (2014). Measurement study of Netflix, Hulu, and a tale of three CDNs. IEEE/ACM Transactions on Networking, 23(6), 1984- 1997. DOI: https://doi.org/10.1109/TNET.2014.2354262

Alhabash, S., Mundel, J., & Hussain, S. A. (2017). Social media advertising. Digital advertising: Theory and research, 285. https://doi.org/10.4324/9781315623252

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives*, *31*(2), 211-36. https://doi.org/10.1257/jep.31.2.211

Amarasingam, A., & Argentino, M. A. (2020). The QAnon conspiracy theory: A security threat in the making. CTC Sentinel, 13(7), 37-44. https://ctc.usma.edu/the-qanon-conspiracy-theory-a-security-threat-in-the-making/

Anspach, N. M., & Carlson, T. N. (2020). What to believe? Social media commentary and belief in misinformation. Political Behavior, 42(3), 697-718. https://doi.org/10.1007/s11109-018-9515-z

BBC News. (2013). Adobe hack: At least 38 million accounts breached. Retrieved March 13, 2021, from https://www.bbc.co.uk/news/technology-24740873

BBC News. (2017). Massive Equifax data breach hits 143 million. Retrieved March 13, 2021, from https://www.bbc.co.uk/news/business-41192163

Benson, V., Saridakis, G. and Tennakoon, H. (2015), "Information disclosure of social media users: Does control over personal information, user awareness and security notices matter?", Information Technology & People, Vol. 28 No. 3, pp. 426- 441. https://doi.org/10.1108/ITP-10-2014-0232

Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General 144, 1 (2015), 114. https://psycnet.apa.org/doi/10.1037/xge0000033

Bhattacharya, R., Devinney, T. M., & Pillutla, M. M. (1998). A formal model of trust based on outcomes. Academy of management review, 23(3), 459-472. https://doi.org/10.5465/amr.1998.926621

Biros, D. P., Daly, M., & Gunsch, G. (2004). The influence of task load and automation trust on deception detection. Group Decision and Negotiation, 13(2), 173-189. https://doi.org/10.1023/B:GRUP.0000021840.85686.57

Blackwell, D., Leaman, C., Tramposch, R., Osborne, C., & Liss, M. (2017). Extraversion, neuroticism, attachment style and fear of missing out as predictors of social media use and addiction. Personality and Individual Differences, 116, 69- 72. https://doi.org/10.1016/j.paid.2017.04.039

Boyd, J. (2002). In community we trust: Online security communication at eBay. *Journal of Computer-Mediated Communication*, *7*(3), JCMC736. https://doi.org/10.1111/j.1083-6101.2002.tb00147.x

Bragg, M., Perkins, W., Sarter, N., Basar, T., Voulgaris, P., Gurbacki, H., Melody, J. and McCray, S., (1998), January. An interdisciplinary approach to inflight aircraft icing safety. In 36th AIAA Aerospace Sciences Meeting and Exhibit (p. 95). https://doi.org/10.2514/6.1998-95

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, *3*(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Brewer, N. T., Gilkey, M. B., Lillie, S. E., Hesse, B. W., & Sheridan, S. L. (2012). Tables or bar graphs? Presenting test results in electronic medical records. *Medical Decision Making*, 32(4), 545-553. https://doi.org/10.1177%2F0272989X12441395

Butler Jr, J. K. (1991). Toward understanding and measuring conditions of trust: Evolution of a conditions of trust inventory. Journal of management, 17(3), 643-663. https://doi.org/10.1177%2F014920639101700307

C. Sievert (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny.*

Chapman and Hall/CRC Florida, 2020. R Package, <URL: https://plotly-r.com>

Cadwalladr, C., & Graham-Harrison, E. (2018). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The guardian*, *17*, 22. from https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election

Celmer, N., Branaghan, R., & Chiou, E. (2018, September). Trust in branded autonomous vehicles & performance expectations: A theoretical framework. In Proceedings of the human factors and ergonomics society annual meeting (Vol. 62, No. 1, pp. 1761-1765). Sage CA: Los Angeles, CA: SAGE Publications. https://doi.org/10.1177/1541931218621398

Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. BMJ Quality & Safety, 28(3), 231-237. http://dx.doi.org/10.1136/bmjqs-2018-008370

Chan, T. H., Jia, K., Gao, S., Lu, J., Zeng, Z., & Ma, Y. (2015). PCANet: A simple deep learning baseline for image classification?. *IEEE transactions on image processing*, 24(12), 5017-5032. https://doi.org/10.1109/TIP.2015.2475625

Chaphalkar, R., & Wu, K. (2020). Students' Reasoning about Variability in Graphs during an Introductory Statistics Course. *International Electronic Journal of Mathematics Education*, 15(2), em0580. https://doi.org/10.29333/iejme/7602

Chavaillaz, A., & Sauer, J. (2017). Operator adaptation to changes in system reliability under adaptable automation. Ergonomics, 60(9), 1261-1272. https://doi.org/10.1080/00140139.2016.1261187

Chavaillaz, A., Schwaninger, A., Michel, S., & Sauer, J. (2019). Expertise, automation and trust in X-ray screening of cabin baggage. Frontiers in psychology, 10, 256. https://doi.org/10.3389/fpsyg.2019.00256

Chavaillaz, A., Schwaninger, A., Michel, S., & Sauer, J. (2020). Some cues are more equal than others: Cue plausibility for false alarms in baggage screening. Applied ergonomics, 82, 102916. https://doi.org/10.1016/j.apergo.2019.102916

Chavaillaz, A., Wastell, D., & Sauer, J. (2016). System reliability, performance and trust in adaptable automation. *Applied ergonomics*, 52, 333-342. https://doi.org/10.1016/j.apergo.2015.07.012

Chen, S. C., & Dhillon, G. S. (2003). Interpreting dimensions of consumer trust in e-commerce. Information technology and management, 4(2), 303-318. https://doi.org/10.1023/A:1022962631249

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., & Su, J. K. (2019). This looks like that: deep learning for interpretable image recognition. In Advances in neural information processing systems (pp. 8930-8941). Doi: https://arxiv.org/abs/1806.10574

Chen, J. Y. C., & Terrence, P. I. (2009). Effects of imperfect automation and individual differences on concurrent performance of military and robotics tasks in a simulated multitasking environment. Ergonomics, 52(8), 907-920. https://doi.org/10.1080/00140130802680773

Chen, J. Y., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. (2014). Situation awareness-based agent transparency. Army Research Lab Aberdeen Proving Ground Md Human Research; Engineering Directorate https://apps.dtic.mil/sti/pdfs/ADA600351.pdf

Culley, K. E., & Madhavan, P. (2013). Trust in automation and automation designers: Implications for HCI and HMI [Editorial]. Computers in Human Behavior, 29(6), 2208–2210. https://doi.org/10.1016/j.chb.2013.04.032

Cummings, M. L. (2017). Automation bias in intelligent time critical decision support systems. In Decision Making in Aviation (pp. 289-294). Routledge. https://doi.org/10.2514/6.2004-6313

Dalsheim, J., & Starrett, G. (2021). Everything Possible and Nothing True: Notes on the Capitol Insurrection. Anthropology Today, 37(2), 26-30. https://doi.org/10.1111/1467-8322.12645

Danks, D., & London, A. J. (2017, August). Algorithmic Bias in Autonomous Systems. *In IJCAI* (pp. 4691-4697). https://dl.acm.org/doi/abs/10.5555/3171837.3171944

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. Management Science, 35(8), 982–1003. Doi: https://doi.org/10.1287/mnsc.35.8.982

Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., & Yanco, H. (2013, March). Impact of robot failures and feedback on real-time trust. In Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction (pp. 251-258). IEEE Press. Doi: https://doi.org/10.1109/HRI.2013.6483596

Desjardins, J. (2018) How Long Does It Take to Hit 50 Million Users? Retrieved September 1, 2018, from http://www.visualcapitalist.com/how-long-does-it-take-to-hit-50-million-users/

Deutsch, M. (1960). The effect of motivational orientation upon trust and suspicion. Human relations, 13(2), 123-139. https://doi.org/10.1177%2F001872676001300202

Doniger, G. M., Foxe, J. J., Schroeder, C. E., Murray, M. M., Higgins, B. A., & Javitt, D. C. (2001). Visual perceptual learning in human object recognition areas: a repetition priming study using high-density electrical mapping. Neuroimage, 13(2), 305-313. Doi: https://doi.org/10.1006/nimg.2000.0684

Duffy, B. R. (2003). Anthropomorphism and the social robot. Robotics and autonomous systems, 42(3-4), 177-190. https://doi.org/10.1016/S0921-8890(02)00374-3

Dunn, J. R., & Schweitzer, M. E. (2005). Feeling and believing: the influence of emotion on trust. Journal of personality and social psychology, 88(5), 736. https://doi.apa.org/doi/10.1037/0022-3514.88.5.736

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, 44(1), 79-94.https://doi.org/10.1518%2F0018720024494856

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. Psychological review, 114(4), 864. https://psycnet.apa.org/doi/10.1037/0033-295X.114.4.864

Furlough, C., Stokes, T., & Gillan, D. J. (2019). Attributing blame to robots: I. The influence of robot autonomy. Human factors https://doi.org/10.1177%2F0018720819880641

Gallagher, K. (2018). Here's a sneak peek at just how big Facebook's trust problem is [exclusive data]. Business Insider. Retrieved December 21, 2018 from http://uk.businessinsider.com/consumers-dont-trust-facebook-at-all-new-survey-data-2018-4?r=US&IR=T

Gefen, D. (2000). E-commerce: the role of familiarity and trust. Omega, 28(6), 725-737. https://doi.org/10.1016/S0305-0483(00)00021-9

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)* (pp. 80- 89). IEEE. Doi: https://arxiv.org/abs/1901.06560

Gilpin, L. H., Testart, C., Fruchter, N., & Adebayo, J. (2019). Explaining explanations to society. arXiv preprint arXiv:1901.06560. https://arxiv.org/abs/1901.06560

Godau, C., Vogelgesang, T., & Gaschler, R. (2016). Perception of bar graphs–A biased impression?. *Computers in Human Behavior*, 59, 67-73. https://doi.org/10.1016/j.chb.2016.01.036

Goddard, K., Roudsari, A., & Wyatt, J. C. (2011). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1), 121-127. https://doi.org/10.1136/amiajnl-2011-000089

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association, 19(1), 121-127. https://doi.org/10.1136/amiajnl-2011-000089

Goddard, K., Roudsari, A., & Wyatt, J. C. (2014). Automation bias: empirical results assessing influencing factors. *International journal of medical informatics*, 83(5), 368-375. https://doi.org/10.1016/j.ijmedinf.2014.01.001

Goëau, H., Joly, A., Yahiaoui, I., Bakić, V., Verroust-Blondet, A., Bonnet, P., Barthélémy, D., Boujemaa, N. and Molino, J.F. (2014). *Plantnet participation at lifeclef2014 plant identification task.* https://hal.archives-ouvertes.fr/halsde-01064569

Gustavo. S. Mesch. 2012. Is online trust and trust in social institutions associated with online disclosure of identifiable information online?. Computers in Human Behavior, 28(4), 1471-1477. https://doi.org/10.1016/j.chb.2012.03.010

Ha, T., Kim, S., Seo, D., & Lee, S. (2020). Effects of explanation types and perceived risk on trust in autonomous vehicles. Transportation research part F: traffic psychology and behaviour, 73, 271-280. https://doi.org/10.1016/j.trf.2020.06.021

Hampton, K. N., Sessions, L. F., & Her, E. J. (2011). Core networks, social isolation, and new media: How Internet and mobile phone use is related to network size and diversity. Information, Communication & Society, 14(1), 130-155. https://doi.org/10.1080/1369118X.2010.513417

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. Human factors, 53(5), 517-527. https://doi.org/10.1177%2F0018720811417254

Hardin, R. (2002). Trust and trustworthiness. Russell Sage Foundation.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *In Advances in psychology* (Vol. 52, pp. 139-183). North-Holland. https://doi.org/10.1016/S0166-4115(08)62386-9

Hawi, N. S., & Samaha, M. (2017). The relations among social media addiction, self-esteem, and life satisfaction in university students. Social Science Computer Review, 35(5), 576-586. https://doi.org/10.1177%2F0894439316660340

Hertz, N., & Wiese, E. (2019). Good advice is beyond all price, but what if it comes from a machine?. *Journal of Experimental Psychology*: Applied. 25(3), 386. http://dx.doi.org/10.1037/xap0000205

Hiar, C. (2021). Twitter bots are a major source of climate disinformation. Retrieved February 22, 2021, from https://www.scientificamerican.com/article/twitter-bots-are -a-major-source-of-climate-disinformation/

Ho, N. T., Sadler, G. G., Hoffmann, L. C., Lyons, J. B., & Johnson, W. W. (2017). Trust of a military automated system in an operational context. Military Psychology, 29(6), 524-541. https://doi.org/10.1037/mil0000189

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3), 407-434 https://doi.org/10.1177%2F0018720814547570

Hosmer, L. T. (1995). Trust: The connecting link between organizational theory and philosophical ethics. Academy of management Review, 20(2), 379-403.

Howard, A. G. (2013). Some improvements on deep convolutional neural network based image classification. *arXiv preprint* arXiv:1312.5402. https://arxiv.org/abs/1312.5402v1

Hussein, A., Elsawah, S., & Abbass, H. A. (2019). Trust Mediating Reliability– Reliance Relationship in Supervisory Control of Human–Swarm Interactions. Human Factors, Doi: https://doi.org/10.1177%2F0018720819879273

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. arXiv preprint arXiv:1602.07360. Doi: https://arxiv.org/abs/1602.07360

Ingram, M., Moreton, R., Gancz, B., & Pollick, F. (2021). Calibrating Trust Toward an Autonomous Image Classifier. Technology, Mind, and Behavior, 2(1). Doi: https://doi.org/10.1037/tmb0000032

Israelsen, B. W., & Ahmed, N. R. (2019). "Dave... I can assure you... that it's going to be all right..." A Definition, Case for, and Survey of Algorithmic Assurances in Human-Autonomy Trust Relationships. *ACM Computing Surveys* (CSUR), 51(6), 1-37. https://doi.org/10.1145/3267338

Jia, X., & Shen, L. (2017). Skin lesion classification using class activation map. arXiv preprint arXiv:1703.01053. Doi: https://arxiv.org/abs/1703.01053

Jin, F., Wang, W., Zhao, L., Dougherty, E., Cao, Y., Lu, C., & Ramakrishnan, N. (2014). Misinformation propagation in the age of twitter. *IEEE Annals of the History of Computing*, *47*(12), 90-94. https://doi.org/10.1109/MC.2014.361

Jing, P., Xu, G., Chen, Y., Shi, Y., & Zhan, F. (2020). The Determinants behind the Acceptance of Autonomous Vehicles: A Systematic Review. *Sustainability*, 12(5), 1719. https://doi.org/10.3390/su12051719

Jones, G. R., & George, J. M. (1998). The experience and evolution of trust: Implications for cooperation and teamwork. Academy of management review, 23(3), 531-546. https://doi.org/10.5465/amr.1998.926625

Jung, A. R. (2017). The influence of perceived ad relevance on social media advertising: An empirical examination of a mediating role of privacy concern. *Computers in Human Behavior*, *70*, 303-309. https://doi.org/10.1016/j.chb.2017.01.008

Kastrenakes, J. (2021). Twitter kept gaining new users after it banned trump. Retrieved February 22, 2021, from https://www.theverge.com/2021/2/9/22275079/twitter -trump-ban-daily-user-growth-q4-2020-earnings

Kay M, Wobbrock J. (2020). *ARTool: Aligned Rank Transform for Nonparametric Factorial ANOVAs*. (URL: https://doi.org/10.5281/zenodo.594511), R package version 0.10.7, <URL: https://github.com/mjskay/ARTool>.

Kim, T., Barasz, K., & John, L. K. (2019). Why am I seeing this ad? The effect of ad transparency on ad effectiveness. *Journal of Consumer Research*, *45*(5), 906-932. https://doi.org/10.1093/jcr/ucy039

Koo, J., Kwac, J., Ju, W., Steinert, M., Leifer, L., & Nass, C. (2015). Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. *International Journal on Interactive Design and Manufacturing* (IJIDeM), 9(4), 269-275. https://doi.org/10.1007/s12008-014-0227-2

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105). https://cs.nju.edu.cn/zhangl/alexnet.pdf

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A. and Duerig, T., (2020). The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 1-26. https://doi.org/10.1007/s11263-020-01316-z

Lanny Lin and Michael A. Goodrich. 2015. Sliding Autonomy for UAV Path-Planning: Adding New Dimensions to Autonomy Management. In Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems (AAMAS '15). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1615–1624. https://www.ifaamas.org/Proceedings/aamas2015/aamas/p1615.pdf

Larson, L., & DeChurch, L. A. (2020). Leading teams in the digital age: Four perspectives on technology and what they mean for leading teams. The Leadership Quarterly, 31(1), 101377. https://doi.org/10.1016/j.leaqua.2019.101377

Lee, D., Hosanagar, K., & Nair, H. S. (2018). Advertising content and consumer engagement on social media: Evidence from Facebook. Management Science, 64(11), 5105 -5131. https://doi.org/10.1287/mnsc.2017.2902

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), 50-80. https://doi.org/10.1518%2Fhfes.46.1.50_30392

Lenth, R. (2020) *emmeans: Estimated Marginal Means, aka Least-Squares Means*, R package version 1.4.8 <URL:https://CRAN.R-project.org/package=emmeans>.

Lewicki, R. J., McAllister, D. J., & Bies, R. J. (1998). Trust and distrust: New relationships and realities. Academy of management Review, 23(3), 438-458. https://doi.org/10.5465/amr.1998.926620

Lewis, D. (2014). iCloud data breach: Hacking and celebrity photos. Retrieved March 13, 2021, from https://www.forbes.com/sites/davelewis/2014/09/02/icloud-data-breach - hacking-and-nude-celebrity-photos/

Limaye, R. J., Sauer, M., Ali, J., Bernstein, J., Wahl, B., Barnhill, A., & Labrique, A. (2020). Building trust while influencing online COVID-19 content in the social media world. The Lancet Digital Health, 2(6), e277-e278. DOI: https://doi.org/10.1016/S2589-7500(20)30084-4

Lipsman, A., Mudd, G., Rich, M., & Bruich, S. (2012). The power of "like": How brands reach (and influence) fans through social-media marketing. Journal of Advertising research, 52(1), 40-52. https://doi.org/10.2501/JAR-52-1-040-052

Lyell, D., & Coiera, E. (2017). Automation bias and verification complexity: a systematic review. Journal of the American Medical Informatics Association, 24(2), 423-431. https://doi.org/10.1093/jamia/ocw105

Lyell, D., Magrabi, F., & Coiera, E. (2018). The effect of cognitive load and task complexity on automation bias in electronic prescribing. *Human Factors*, 60(7), 1008-1021. https://doi.org/10.1177%2F0018720818781224

Lyons, J. B., & Guznov, S. Y. (2019). Individual differences in human–machine trust: A multi-study look at the perfect automation schema. Theoretical Issues in Ergonomics Science, 20(4), 440-458. https://doi.org/10.1080/1463922X.2018.1491071

Lyons, J. B., & Stokes, C. K. (2012). Human–human reliance in the context of automation. Human factors, 54(1), 112-121. https://doi.org/10.1177%2F0018720811427034

Lyons, J. B., Ho, N. T., Fergueson, W. E., Sadler, G. G., Cals, S. D., Richardson, C. E., & Wilkins, M. A. (2016). Trust of an automatic ground collision avoidance technology: A fighter pilot perspective. Military Psychology, 28(4), 271-277. https://doi.org/10.1037/mil0000124

Machackova, H., Cerna, A., Sevcikova, A., Dedkova, L., & Daneback, K. (2013). Effectiveness of coping strategies for victims of cyberbullying. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *7*(3). https://doi.org/10.5817/CP2013-3-5

Madhavan, P., & Wiegmann, D.A., (2007) Similarities and differences between human–human and human–automation trust: an integrative review, Theoretical Issues in Ergonomics Science, 8:4, 277-301, DOI: https://doi.org/10.1080/14639220500337708

MathWorks (2019). Investigate Network Predictions Using Class Activation Mapping. Retrieved from https://www.mathworks.com/help/deeplearning/ug/investigate- network-predictions-using-class-activation-mapping.html

MATLAB. (2017). version 9.2 (R2017a). Natick, Massachusetts: The MathWorks Inc.

Maurtua, I., Ibarguren, A., Kildal, J., Susperregi, L., & Sierra, B. (2017). Human–robot collaboration in industrial applications: Safety, interaction and trust. International Journal of Advanced Robotic Systems, 14(4). https://doi.org/10.1177%2F1729881417716010

Mazurek, M. O. (2013). Social media use among adults with autism spectrum disorders. Computers in Human Behavior, 29(4), 1709-1714. https://doi.org/10.1016/j.chb.2013.02.004

McGuirl, J. M., & Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. Human factors, 48(4), 656-665. https://doi.org/10.1518%2F001872006779166334

McGuirl, J., Sarter, N., & Woods, D. (2009). Effects of real-time imaging on decision-making in a simulated incident command task. International Journal of Information Systems for Crisis Response and Management (IJISCRAM), 1(1), 54-69. https://doi.org/10.4018/jiscrm.2009010105

McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.C., Darzi, A. and Etemadi, M., (2020). International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788), pp.89-94. https://doi.org/10.1038/s41586-019-1799-6

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. Information systems research, 13(3), 334-359. https://doi.org/10.1287/isre.13.3.334.81

Mercado, J. E., Rupp, M. A., Chen, J. Y., Barnes, M. J., Barber, D., & Procci, K. (2016). Intelligent agent transparency in human–agent teaming for Multi-UxV management. Human factors, 58(3), 401-415. https://doi.org/10.1177/0018720815621206

Merritt, S. M. (2011). Affective processes in human–automation interactions. *Human Factors*, 53(4), 356-370. https://doi.org/10.1177%2F0018720811411912

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I trust it, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human factors*, 55(3), 520-534. https://doi.org/10.1177%2F0018720812465081

Merritt, S. M., Lee, D., Unnerstall, J. L., & Huber, K. (2015). Are Well-Calibrated Users Effective Users? Associations Between Calibration of Trust and Performance on an Automation-Aided Task. Human Factors, 57(1), 34–47. https://doi.org/10.1177/0018720814561675

Mitchell, R. W., Thompson, N. S., & Miles, H. L. (Eds.). (1997). Anthropomorphism, anecdotes, and animals. Suny Press.

Morgan-Thomas, A., & Veloutsou, C. (2013). Beyond technology acceptance: Brand relationships and online brand experience. Journal of Business Research, 66(1), 21-27. https://doi.org/10.1016/j.jbusres.2011.07.019

Morrar, R., Arman, H., & Mousa, S. (2017). The fourth industrial revolution (Industry 4.0): A social innovation perspective. Technology Innovation Management Review, 7(11), 12-20.

https://timreview.ca/sites/default/files/Issue_PDF/TIMReview_November2017.pdf#page=12

Mosier, K. L., Skitka, L. J., Heers, S., & Burdick, M. (2017). Automation bias: Decision making and performance in high-tech cockpits. In Decision Making in Aviation (pp. 271-288). Routledge. https://doi.org/10.1207/s15327108ijap0801_3

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, 27(5-6), 527-539. https://doi.org/10.1016/S0020-7373(87)80013-5

Muir, B.M., (1994) Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems, ERGONOMICS, 37:11, 1905-1922, DOI: https://doi.org/10.1080/00140139408964957

Mukhopadhyay, A., Mukherjee, I., & Biswas, P. (2020, September). Decoding CNN based Object Classifier Using Visualization. In 12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (pp. 50-53). Doi: https://doi.org/10.1145/3409251.3411721

Murphy, M. (2017, December 18). *Artificial intelligence will detect child abuse images to save police from trauma*. Retrieved November 1, 2019, from: https://www.telegraph.co.uk/technology/2017/12/18/artificial-intelligence-will-detect-child-abuse-images-save/

Neurath, P. W., Brand, D. H., & Schreiner, E. D. (1969). MAN-MACHINE INTERACTION FOR IMAGE PROCESSING. Annals of the New York Academy of Sciences, 157(1), 324-338. https://doi.org/10.1111/j.1749-6632.1969.tb12669.x

Niu, D., Terken, J., & Eggen, B. (2018). Anthropomorphizing information to enhance trust in autonomous vehicles. Human Factors and Ergonomics in Manufacturing & Service Industries, 28(6), 352-359. https://doi.org/10.1002/hfm.20745

Oyeyemi, S. O., Gabarron, E., & Wynn, R. (2014). Ebola, Twitter, and Misinformation: A Dangerous Combination?. Bmj, 349. https://doi.org/10.1136/bmj.g6178

Ozdemir, B., & Kumral, M. (2019). Analysing human effect on the reliability of mining equipment. International Journal of Heavy Vehicle Systems, 26(6), 872-887. Doi: https://doi.org/10.1504/IJHVS.2019.102716

Özlen, M. K., & Djedovic, I. (2017). Online banking acceptance: The influence of perceived system security on perceived system quality. Accounting and Management Information Systems, 16(1), 164-178. https://econpapers.repec.org/RePEc:ami:journl:v:16:y:2017:i:1:p:164-178

Pak, R., Rovira, E., McLaughlin, A. C., & Baldwin, N. (2017). Does the domain of technology impact user trust? Investigating trust in automation across different consumer-oriented domains in young adults, military, and older adults. Theoretical issues in ergonomics science, 18(3), 199-220. https://doi.org/10.1080/1463922X.2016.1175523

Pammer, K., Gauld, C., McKerral, A., & Reeves, C. (2021). “They have to be better than human drivers!” Motorcyclists’ and cyclists’ perceptions of autonomous vehicles.

Transportation research part F: traffic psychology and behaviour, 78, 246-258. https://doi.org/10.1016/j.trf.2021.02.009

Papenmeier, A., Englebienne, G., & Seifert, C. (2019). How model accuracy and explanation fidelity influence user trust. arXiv preprint arXiv:1907.12652. Doi: https://arxiv.org/abs/1907.12652

Parasuraman, R., & Miller, C. A. (2004). Trust and etiquette in high-criticality automated systems. Communications of the ACM, 47(4), 51-55. https://doi.org/10.1145/975817.975844

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2), 230-253. https://doi.org/10.1518%2F001872097778543886

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, 30(3), 286-297. https://doi.org/10.1109/3468.844354

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2008). Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. Journal of cognitive engineering and decision making, 2(2), 140-160. https://doi.org/10.1518%2F155534308X284417

Pariser, E. (2011). The filter bubble: What the Internet is hiding from you. Penguin UK.

Park, K. G., Han, S., & Kaid, L. L. (2013). Does social networking service usage mediate the association between smartphone usage and social capital?. New media & society, 15(7), 1077-1093. https://doi.org/10.1177%2F1461444812465927

Parris, L., Lannin, D. G., Hynes, K., & Yazedjian, A. (2020). Exploring social media rumination: associations with bullying, cyberbullying, and distress. Journal of interpersonal violence https://doi.org/10.1177/0886260520946826

Patro, B. N., Lunayach, M., Patel, S., & Namboodiri, V. P. (2019). U-cam: Visual explanation using uncertainty based class activation maps. In Proceedings of the IEEE International Conference on Computer Vision (pp. 7444-7453). Doi: https://doi.org/10.1109/ICCV.2019.00754

Perkins, L., Miller, J. E., Hashemi, A., & Burns, G. (2010, September). Designing for human-centered systems: Situational risk as a factor of trust in automation. In Proceedings of the human factors and ergonomics society annual meeting (Vol. 54, No. 25, pp. 2130-2134). Sage CA: Los Angeles, CA: SAGE Publications. https://doi.org/10.1177%2F154193121005402502

Pittman, M., & Reich, B. (2016). Social media and loneliness: Why an Instagram picture may be worth more than a thousand Twitter words. *Computers in Human Behavior*, *62*, 155-167. https://doi.org/10.1016/j.chb.2016.03.084

R Core Team (2020). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.<URL: https://www.R-project.org/>.

Rainie, L. (2018). "How Americans Feel about Social Media and Privacy." Pew Research Center. March 27, 2018. Retrieved: December 26, 2018. http://www.pewresearch.org/fact-tank/2018/03/27/americans-complicated-feelings-about-social-media-in-an-era-of-privacy-concerns/.

Read, W. H. (1962). Upward communication in industrial hierarchies. Human relations, 15(1), 3-15. https://doi.org/10.1177%2F001872676201500101

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). Why should I trust you?: Explaining the predictions of any classifier. *In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144). ACM. https://doi.org/10.1145/2939672.2939778

Ribera, M., & Lapedriza, A. (2019, March). Can we do better explanations? A proposal of user-centered explainable AI. In IUI Workshops. Doi: http://hdl.handle.net/10609/99643

Rice, S., Clayton, K., & McCarley, J. (2017). The effects of automation bias on operator compliance and reliance. In Human factors issues in combat identification (pp. 265-276). CRC Press. https://doi.org/10.1201/9781315587387

Roberts, J. A., & David, M. E. (2020). The social media party: Fear of missing out (FoMO), social media intensity, connection, and well-being. *International Journal of Human–Computer Interaction*, *36*(4), 386-392. https://doi.org/10.1080/10447318.2019.1646517

Rogers, H., Khasawneh, A., Bertrand, J., & Chalil, K. (2019, November). Understanding reliance and trust in decision aids for UAV target identification. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 63, No. 1, pp. 1953-1954). Sage CA: Los Angeles, CA: SAGE Publications. https://doi.org/10.1177%2F1071181319631172

Rotter, J. B. (1980). Interpersonal trust, trustworthiness, and gullibility. American psychologist, 35(1), 1. https://psycnet.apa.org/doi/10.1037/0003-066X.35.1.1

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. Academy of management review, 23(3), 393-404. https://doi.org/10.5465/amr.1998.926617

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x

Ryan, T. J., Alarcon, G. M., Walter, C., Gamble, R., Jessup, S. A., Capiola, A., & Pfahler, M. D. (2019, July). Trust in automated software repair. *In International Conference on Human-Computer Interaction* (pp. 452-470). Springer, Cham. https://doi.org/10.1007/978-3-030-22351-9_31

Sabella, R. A., Patchin, J. W., & Hinduja, S. (2013). Cyberbullying myths and realities. *Computers in Human behavior*, *29*(6), 2703-2711. https://doi.org/10.1016/j.chb.2013.06.040

Satterfield, K., Baldwin, C., de Visser, E., & Shaw, T. (2017, September). The influence of risky conditions in trust in autonomous systems. In Proceedings of the Human Factors and

Ergonomics Society Annual Meeting (Vol. 61, No. 1, pp. 324-328). Sage CA: Los Angeles, CA: SAGE Publications. https://doi.org/10.1177%2F1541931213601562

Sauer, J., & Chavaillaz, A. (2017). The use of adaptable automation: Effects of extended skill lay-off and changes in system reliability. Applied Ergonomics, 58, 471-481. Doi: https://doi.org/10.1016/j.apergo.2016.08.007

Sauer, J., Chavaillaz, A., & Wastell, D. (2016). Experience of automation failures in training: effects on trust, automation bias, complacency and performance. *Ergonomics*, 59(6), 767-780. https://doi.org/10.1080/00140139.2015.1094577

Schwark, J., Dolgov, I., Graves, W., & Hor, D. (2010, September). The influence of perceived task difficulty and importance on automation use. I*n Proceedings of the and Ergonomics Society Annual Meeting* (Vol. 54, No. 19, pp. 1503-1507). Sage CA: Los Angeles, CA: SAGE Publications. https://doi.org/10.1177%2F154193121005401931

Selkowitz, A. R., Larios, C. A., Lakhmani, S. G., & Chen, J. Y. (2017). Displaying information to support transparency for autonomous platforms. *In Advances in Human Factors in Robots and Unmanned Systems* (pp. 161-173). Springer, Cham. https://doi.org/10.1007/978-3-319-41959-6_14

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad- cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626). Doi: https://doi.org/10.3389/frai.2021.703504

Sheridan, T. B., & Verplank, W. L. (1978). Human and computer control of undersea teleoperators. Massachusetts Inst of Tech Cambridge Man-Machine Systems Lab. https://apps.dtic.mil/sti/pdfs/ADA057655.pdf

Solon, O. (2018). "Facebook Stocks Plummet More than 20% amid Concerns over Growth." The Guardian. July 25, 2018. Retrieved December 28, 2018 from: https://www.theguardian.com/technology/2018/jul/25/facebook-stocks-second-quarter-revenue-user-growth

Sterrett, D., Malato, D., Benz, J., Kantor, L., Tompson, T., Rosenstiel, T., Sonderman, J. and Loker, K., 2019. Who shared it?: Deciding what news to trust on social media. *Digital Journalism*, *7*(6), pp.783-801. DOI: https://doi.org/10.1080/21670811.2019.1623702

Sujan, M., Furniss, D., Grundy, K., Grundy, H., Nelson, D., Elliott, M., White, S., Habli, I. and Reynolds, N., 2019. Human factors challenges for the safe use of artificial intelligence in patient care. BMJ health & care informatics, 26(1). https://doi.org/10.1136%2Fbmjhci-2019-100081

Sung, A., & Klein, D. D. (2021). January 6th and President Trump: A Study of Social Media in Today's America. Proceedings of the 23rd International RAIS Conference on Social Sciences and Humanities http://rais.education/wp-content/uploads/2021/09/0095.pdf

Thüring, M., & Mahlke, S. (2007). Usability, aesthetics and emotions in human–technology interaction. International journal of psychology, 42(4), 253-264. https://doi.org/10.1080/00207590701396674

Tomlinson, E. C., & Mryer, R. C. (2009). The role of causal attribution dimensions in trust repair. Academy of Management Review, 34(1), 85-104. https://doi.org/10.5465/amr.2009.35713291

Tomlinson, E. C., Dineen, B. R., & Lewicki, R. J. (2004). The road to reconciliation: Antecedents of victim willingness to reconcile following a broken promise. Journal of management, 30(2), 165-187. https://doi.org/10.1016%2Fj.jm.2003.01.003

Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., Pearson, G. and Kaplan, L., (2020). Rapid trust calibration through interpretable and uncertainty-aware AI. *Patterns*, 1(4), p.100049. https://doi.org/10.1016/j.patter.2020.100049

Tufekci, Z. (2015). Algorithmic harms beyond Facebook and Google: Emergent challenges of computational agency. *Colo. Tech. LJ*, *13*, 203. http://ai.ethicsworkshop.org/Library/LibContentAcademic/Tufekci-AlgorithmsAgency.pdf

Tulek, Z., & Arnell, L. (2019). Facebook Eavesdropping Through the Microphone for Marketing Purpose (Dissertation). Retrieved from: http://urn.kb.se/resolve?urn=urn:nbn:se:bth-18232

Venkatesh, V. (2015). Technology acceptance model and the unified theory of acceptance and use of technology. Wiley Encyclopedia of Management, 7, 1–9. Doi: https://doi.org/10.1002/9781118785317.weom070047

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. Management science, 46(2), 186-204. Doi: https://doi.org/10.1287/mnsc.46.2.186.11926

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. MIS quarterly, 425-478. https://doi.org/10.2307/30036540

Venkatesh, V., Thong, J. Y., & Xu, X. (2012). Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. *MIS quarterly*, 36(1), 157-178. https://www.jstor.org/stable/41410412

Verame, J. K. M., Costanza, E., & Ramchurn, S. D. (2016, May). The effect of displaying system confidence information on the usage of autonomous systems for non-specialist applications: A lab study. In Proceedings of the 2016 chi conference on human factors in computing systems (pp. 4908-4920). ACM https://doi.org/10.1145/2858036.2858369

Verberne, F. M., Ham, J., & Midden, C. J. (2012). Trust in smart systems: Sharing driving goals and giving information to increase trustworthiness and acceptability of smart systems in cars. Human factors, 54(5), 799-810. https://doi.org/10.1177%2F0018720812443825

Verberne, F. M., Ham, J., & Midden, C. J. (2015). Trusting a virtual driver that looks, acts, and thinks like you. Human factors, 57(5), 895-909. https://doi.org/10.1177%2F0018720815580749

de Visser, E. J., Cohen, M., Freedy, A., & Parasuraman, R. (2014, June). A design methodology for trust cue calibration in cognitive agents. In International conference on virtual, augmented and mixed reality (pp. 251-262). Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-319-07458-0_24

de Visser, E. J., Krueger, F., McKnight, P., Scheid, S., Smith, M., Chalk, S., & Parasuraman, R. (2012, September). The world is not enough: Trust in cognitive agents. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 56, No. 1, pp. 263-267). Sage CA: Los Angeles, CA: Sage Publications. https://doi.org/10.1177%2F1071181312561062

de Visser, E.J., Monfort, S.S., Goodyear, K., Lu, L., O'Hara, M., Lee, M.R., Parasuraman, R. and Krueger, F., (2017). A little anthropomorphism goes a long way: Effects of oxytocin on trust, compliance, and team performance with automated agents. Human factors, 59(1), pp.116-133. https://doi.org/10.1177%2F0018720816687205

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, 22(3), 331. https://psycnet.apa.org/doi/10.1037/xap0000092

de Visser, E. J., Pak, R., & Shaw, T.H., (2018): From 'automation' to 'autonomy': the importance of trust repair in human–machine interaction, Ergonomics, DOI: https://doi.org/10.1080/00140139.2018.1457725

de Visser, E., & Parasuraman, R. (2011). Adaptive aiding of human-robot teaming: Effects of imperfect automation on performance, trust, and workload. Journal of Cognitive Engineering and Decision Making, 5(2), 209-231. Doi: https://doi.org/10.1177%2F1555343411410160

Warner-Søderholm, G., Bertsch, A., Sawe, E., Lee, D., Wolfe, T., Meyer, J., Engel, J. and Fatilua, U.N., 2018. Who trusts social media?. Computers in human behavior, 81, pp.303-315. https://doi.org/10.1016/j.chb.2017.12.026

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. Journal of Experimental Social Psychology, 52, 113-117. https://doi.org/10.1016/j.jesp.2014.01.005

West, L. A. (1967). An Agricultural Machinery Museum. Agricultural History, 41(3), 267-274. https://www.jstor.org/stable/3740340

Wickham, H. (2016). ggplot2: elegant graphics for data analysis. Springer. <URL: https://ggplot2.tidyverse.org>.

Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011, May). The aligned rank transform for nonparametric factorial analyses using only anova procedures. *In Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 143-146). https://doi.org/10.1145/1978942.1978963

Woods, H. C., & Scott, H. (2016). #Sleepyteens: Social media use in adolescence is associated with poor sleep quality, anxiety, depression and low self-esteem. Journal of adolescence, 51, 41-49. https://doi.org/10.1016/j.adolescence.2016.05.008

Wright, J. L., Chen, J. Y., & Lakhmani, S. G. (2019). Agent transparency and reliability in human–robot interaction: the influence on user confidence and perceived reliability. IEEE Transactions on Human-Machine Systems, 50(3), 254-263. Doi: https://doi.org/10.1109/THMS.2019.2925717

Wright, J. L., Chen, J. Y., Lakhmani, S., & Selkowitz, A. (2020). Agent transparency for an autonomous squad member: depth of reasoning and reliability. DEVCOM Army Research Laboratory Aberdeen Proving Ground United States. https://apps.dtic.mil/sti/pdfs/AD1116568.pdf

Wright, T. J., Horrey, W. J., Lesch, M. F., & Rahman, M. M. (2016). Drivers' trust in an autonomous system: Exploring a covert video-based measure of trust. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 60(1), 1334–1338. https://doi.org/10.1177/1541931213601308

Wu, W. Y., Quyen, P. T. P., & Rivas, A. A. A. (2017). How e-servicescapes affect customer online shopping intention: the moderating effects of gender and online purchasing experience. Information Systems and e-Business Management, 15(3), 689-715. https://doi.org/10.1007/s10257-016-0323-x

Xiao, X., Borah, P., & Su, Y. (2021). The dangers of blind trust: Examining the interplay among social media news use, misinformation identification, and news trust on conspiracy beliefs. Public Understanding of Science, 30(8), 977-992. https://doi.org/10.1177%2F0963662521998025

Yang, W., Huang, H., Zhang, Z., Chen, X., Huang, K., & Zhang, S. (2019). Towards rich feature discovery with class activation maps augmentation for person re- identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1389-1398). Doi: https://doi.ieeecomputersociety.org/10.1109/CVPR.2019.00148

Yu, K., Berkovsky, S., Taib, R., Zhou, J., & Chen, F. (2019, March). Do I trust my machine teammate? An investigation from perception to decision. *In Proceedings of the 24th International Conference on Intelligent User Interfaces* (pp. 460-468). https://doi.org/10.1145/3301275.3302277

Zhang, Y., Liao, Q. V., & Bellamy, R. K. (2020, January). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*(pp.295-305). https://doi.org/10.1145/3351095.3372852