Ergonomics

ISSN: 0014-0139 (Print) 1366-5847 (Online) Journal homepage: http://www.tandfonline.com/loi/terg20

From 'automation' to 'autonomy': the importance of trust repair in human–machine interaction

Ewart J. de Visser, Richard Pak & Tyler H. Shaw

To cite this article: Ewart J. de Visser, Richard Pak & Tyler H. Shaw (2018): From 'automation' to 'autonomy': the importance of trust repair in human–machine interaction, Ergonomics, DOI: 10.1080/00140139.2018.1457725

To link to this article: https://doi.org/10.1080/00140139.2018.1457725

Accepted author version posted online: 26 Mar 2018.
Published online: 09 Apr 2018.

ERGONOMICS, 2018
https://doi.org/10.1080/00140139.2018.1457725

Taylor & Francis
Taylor & Francis Group

From 'automation' to 'autonomy': the importance of trust repair in human–machine interaction

Ewart J. de Visser[a,b], Richard Pak[c] and Tyler H. Shaw[a]

[a]Human Factors and Applied Cognition, Department of Psychology, George Mason University, Fairfax, VA, USA; [b]Warfighter Effectiveness Research Center, Department of Behavioral Sciences and Leadership, United States Air Force Academy, Colorado Springs, CO, USA; [c]Department of Psychology, Clemson University, Clemson, SC, USA

ABSTRACT

Modern interactions with technology are increasingly moving away from simple human use of computers as tools to the establishment of human *relationships* with autonomous entities that carry out actions on our behalf. In a recent commentary, Peter Hancock issued a stark warning to the field of human factors that attention must be focused on the appropriate design of a new class of technology: highly autonomous systems. In this article, we heed the warning and propose a human-centred approach directly aimed at ensuring that future human–autonomy interactions remain focused on the user's needs and preferences. By adapting literature from industrial psychology, we propose a framework to infuse a unique human-like ability, building and actively repairing trust, into autonomous systems. We conclude by proposing a model to guide the design of future autonomy and a research agenda to explore current challenges in repairing trust between humans and autonomous systems.

Practitioner Summary: This paper is a call to practitioners to re-cast our connection to technology as akin to a relationship between two humans rather than between a human and their tools. To that end, designing autonomy with trust repair abilities will ensure future technology maintains and repairs relationships with their human partners.

ARTICLE HISTORY
Received 16 June 2017
Accepted 15 March 2018

KEYWORDS
Trust repair; autonomy; automation; humanness; human–machine teaming

1. Introduction

In a recent commentary, Hancock (2017) issued a stark warning to the field of human factors that attention must be focused on the appropriate design of a new class of technology: highly autonomous systems. In that warning, he argued that society was in the midst of a major change with the introduction of more independent, autonomous systems that seem to be of a different class from conventional automation. He further argued that while development of these new systems is proceeding at breakneck speed by practitioners, the study of the psychological and human factors implications might be ignored. In this article, we heed the warning and propose a human-centred approach directly aimed at providing guidance to the designers of these future systems. Finally, to further the discussion regarding these new autonomous systems (Endsley 2017a; Kaber 2017a; Sheridan 2016; Woods 2016).

1.1. New forms of automation and autonomy redefine our relationship with technology

Modern interactions with technology are increasingly moving away from simple human use of computers as tools to the establishment of human *relationships* with autonomous entities. In contrast to a conventional automated system designed to carry out a limited set of pre-programmed supervised tasks on behalf of the user, autonomy is technology (either hardware or software) designed to carry out a user's goals, but that does not require supervision. Recent examples of such highly autonomous technology is the Stuxnet virus and its ability to plan independently and then automatically spread all over the internet (Kushner 2013) or the Mirai Botnet, that performed the most sophisticated distributed-denial-of-service attacks to date (Graff 2017). It is also important to consider that highly autonomous systems are projected to be integrated into our day-to-day lives sooner rather than later. For example, personnel at the highest reaches of the Air Force have been discussing an initiative entitled the 'Loyal Wingman', which involves autonomous aircraft flying alongside manned flight lead aircraft (e.g. Gurney 2013; Humphreys et al. 2015). Moreover, 'self-driving' vehicles are already beginning to propagate through our society, with companies like Google (Brown 2011) launching tests of autonomous vehicles, as well as prototype autonomous public modes of transportation (e.g. busses) being tested in Switzerland

CONTACT Ewart J. de Visser edevisse@gmu.edu

and Finland (Meyer et al. 2017; Roemer et al. 2017; Salonen 2018). For these reasons, it is critically important that we begin investigating the issues associated with the use of autonomous systems.

Systems that exhibit autonomy are distinguished by their capability to learn and change over time, dynamically setting its own goals, and the ability to adapt to local conditions via external sensor information or updated input data. Designers may preside over the start state and parameters of such systems, but once deployed, autonomy will evolve with use in different environments. This means that this technology has the potential to evolve in unexpected ways (Kurzweil 2005). Because of the potential unpredictability of these systems, it may also be more likely that autonomy will surprise human partners to an even greater extent than simple automated systems (Sarter and Woods 1997), and this has the potential to greatly affect trust and adoption of such technology. Moreover, because autonomous systems are less like tools and more like assistants or collaborators, they will likely spend more time with us because of their independence and this will lead to a longer term relationship with this technology. For example, while the new classes of always-on, always-near virtual assistants such as the Amazon Echo, Alexa or Siri do not yet rise to the level of fully autonomous systems, they are designed to reside in our home permanently or remain on-the-body. Finally, these virtual assistants are designed to learn our behaviours and preferences over time, gather contextual information from the environment (e.g. location), and learn from our constantly updated user profiles. As a result of these more capable systems, our attachments and motivations towards these technologies will likely grow deeper and more intimate (Szalma 2014; de Visser, Monfort, et al., 2017; Wiese et al. 2017).

Because of some of the previously stated differences between automation and autonomous systems, we argue that a fundamentally different approach to enhancing human–machine interaction is necessary; one where the relationship between autonomy and the human is constantly changing, reflecting the adaptivity of autonomy itself and human perceptions. Instead, the paradigm of human–autonomy interaction should emulate the rich interactions of relationships between people and should adopt human–human models as their initial standards; the autonomy should be able to take advantage of extant human capabilities of detecting incidental information from others. With that in mind, and inspired by models of human–human interaction, we propose that future design of autonomy ought to take cues from the social sciences. Although we believe this a prudent general method for approaching a new design approach to autonomy, we focus in this article on one specific human relationship trait that we believe to be beneficial to the human–autonomy relationship. To that end, we outline one possible mechanism that serves as a fundamental aspect of healthy relationships: trust repair. Specifically, we suggest that researchers must (1) re-cast their view of the human–autonomy system as beyond simple 'interactions' and more as a 'relationship', and (2) further extend the application of research on human–human relationships to human–autonomy (Madhavan and Wiegmann 2004, 2007a).

1.2. Further distinguishing automation from autonomy

While we are in agreement with Hancock's initial observation of the differences between automation and autonomy (2017), we build upon this idea to make the distinction even clearer and to project the implications of this difference into a future research agenda. For the purposes of making recommendations for the design of autonomous systems to optimise human–autonomy interaction, we outlined a framework that juxtaposes the previously defined concept of *autonomy* with the new design concept of *humanness design*, inspired by a social psychological construct (Haslam 2006; Haslam et al. 2005). We define *humanness design* as any strategy, mechanism, and feature of the autonomy designed to connect and communicate with a human. Our concept of humanness is intentionally broad and defined simply to capture the design intent to connect and communicate with other humans, but our concept encompasses Haslam's (2006) two senses of humanness that includes (1) uniquely human characteristics; human features that distinguish us from other living organisms (i.e. civility, refinement, moral sensibility, rationality, maturity), and (2) human nature; human features that represent the essence of being human (i.e. emotional responsiveness, interpersonal warmth, cognitive openness, agency, depth). Arguably, any autonomous design will require at least some humanness (i.e. a mechanism to accept user input), but higher levels on this axis approximate the appearance, emotional and behavioural and capabilities of humans.

Figure 1 illustrates how we might build upon this existing definition of autonomous systems. The figure illustrates a conceptual representation space of level of autonomy (i.e. ability to act unsupervised) against humanness design (i.e. features, strategies and mechanisms designed to connect and communicate with humans). Using these two dimensions, we created four distinct quadrants that represent seemingly qualitative different classes of machines. The lower left quadrant represents current or near-future, and fictional representations, meant to represent exemplars of autonomous systems. For example, machines designed to play strategy games such as Go or Chess

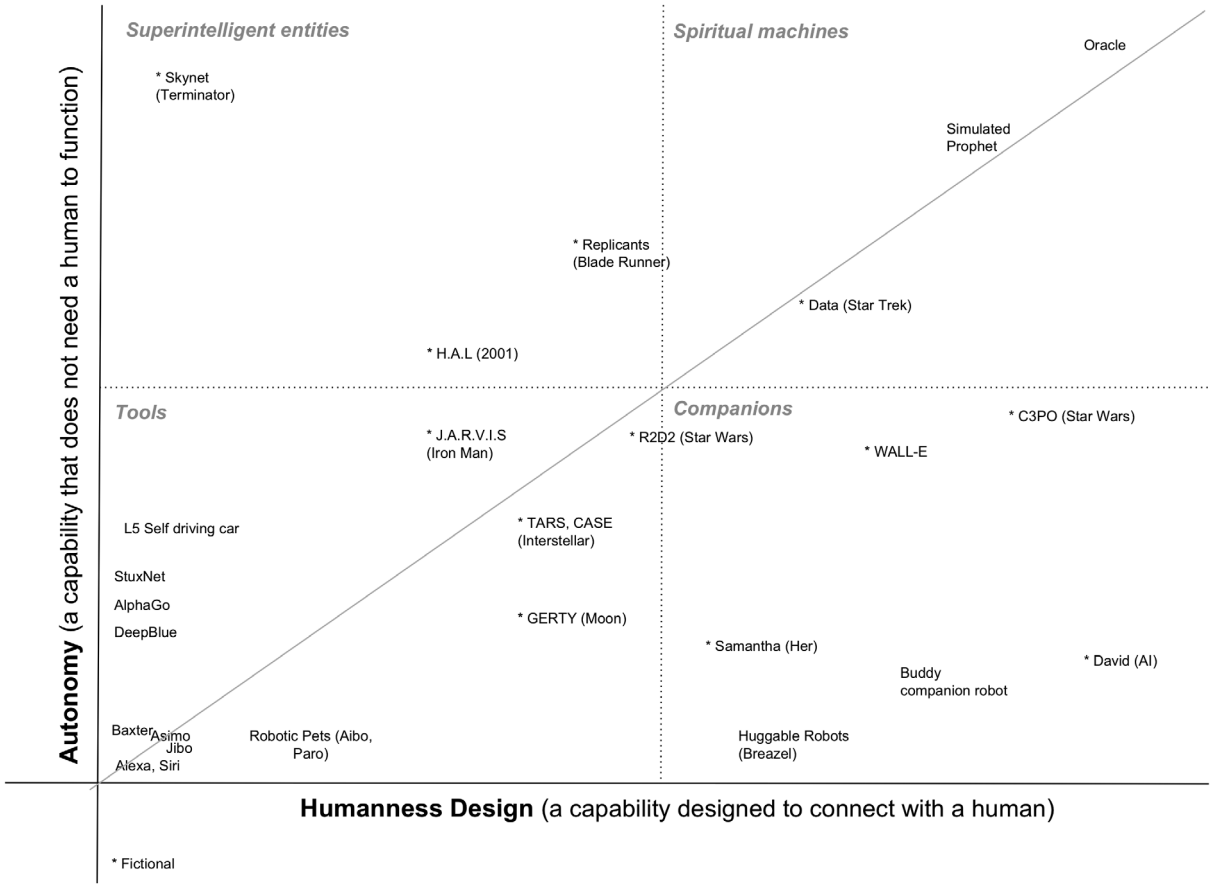

Figure 1. Conceptual representation of existing and fictional agents plotted with degrees of autonomy and humanness design.

(AlphaGo, DeepBlue) represent a moderate level of autonomy because these systems operate toward a goal, such as winning the game, and require no supervision. These forms of autonomy are, however, very limited in humanness because they are not explicitly designed to communicate or interface directly with a user; the machines are only algorithms designed for a singular purpose. Within the same quadrant are current robotic pets such as Aibo or Paro, which are explicitly designed to interact and communicate with humans, but have little to no autonomy and may have a diffuse goal or no goal other than to provide entertainment, companionship or comfort to the human. Higher levels of autonomy or humanness design are represented by mostly fictional examples. As such, the relative positions within quadrants are meant to be approximate and is based on their portrayal in the film. Thus, small distances in the figure should not be construed as precision.

High autonomy, but moderate humanness (upper left quadrant) represents tools that act on their own towards a goal but where communication with human operators is not a primary concern and so they may have limited abilities to communicate (e.g. rudimentary modes of input and output). Perhaps the classic example is of the super-intelligent AI Skynet, which in the Terminator movie franchise was highly autonomous (i.e. it had its own goal of self preservation and acted on those goals autonomously), but had limited capabilities or consideration for direct human communication. High humanness, but moderate autonomy (lower right) represents machines that are explicitly designed to communicate and interact with humans so they may have highly developed modes of input and output (natural voice, gestural communications, humour or attitude) or use appearance cues (anthropomorphic appearance) designed to blend in with humans. However, they may have little ability to generate or act on their own goals or that do not have goals other than to provide companionship. Technologies with high humanness and high autonomy (upper right quadrant) represent machines that have their own intentions and goals and operate with virtually no human oversight.

The framework both captures current, and anticipated but fictional technologies. From this notional framework, it is possible to draw several conclusions.

(1) *Increasing humanness is desired if the design requires a connection and communication with a human user.* We believe design with more humanness is necessary with the increasing complexity

and potential for mismatch between user perception and capability of AI (Semigran et al. 2016). Humanness will become a required interface feature because we believe it represents an optimal (high-bandwidth, but low-resource-demanding) way to convey the complexities of autonomous behaviour to humans (e.g. facial expressions; Chernoff 1973; Nass and Lee 2001; Nelson 2007; Semigran et al. 2016).

(2) *Including humanness into designs may be a balancing factor to prevent undesirable autonomy.* To create a balanced middle of the road approach for autonomy, we advocate autonomy that incorporates humanness early on. We believe that technology should be created for a human world. Autonomy without humanness will create powerful machines, but these machines will be disconnected from humanity (the least of which is the out of the loop phenomenon). Humanness without autonomy creates congenial social machines, but these machines are the equivalent of our former tools. The combination of both design aspects may productively enhance autonomy (Waytz, Heafner, and Epley 2014).

(3) *Balanced autonomy-humanness may represent a way to prevent unfriendly entities from controlling humanity.* The goals and intent of an autonomous system must be made transparent to the user. Examples of subtle unexpected and undesired effects are biases in algorithms (Facebook, Amazon, Google) and a company imposing its commercial goals onto the user such as gambling (Riva, Sacchi, and Brambilla 2015).

For all of the previously stated differences between automated systems and systems that exhibit autonomy, previously identified problems with automation (e.g. out of the loop syndrome, mode awareness, complacency, trust) are expected to not only exist, but be magnified, a phenomenon known as the lumberjack effect (Onnasch et al. 2014). In addition, we expect a new class of problems unique to systems with autonomy; those brought about by the independence of these systems, but also by the ways in which they will interact with users (Clare, Cummings, and Repenning 2015). For example, while conventional automation research has been highly focused on the effects of automation on user performance and to some extent subjective perceptions, we expect that human interactions with highly autonomous systems will result in highly emotional reactions, and that the acceptance and usage of autonomy will be dominated by social and psychological factors to a greater extent than with conventional automation.

Another primary challenge posed by autonomous systems concerns the level of information the system should convey to the human operator. The human factors literature has seen a shift in the use of terminology regarding the relation between humans and machines, with researchers now referring to 'human-machine teaming (HMT)' in place of the more traditional 'human-computer interaction (HCI) (Chen and Barnes 2014)'. This change in terminology represents a change in the underlying HCI framework, such that machines are evolving from 'tools' (i.e. automation) to 'teammates' (i.e. autonomy). In human–human partnerships, communication has always been viewed as a vital aspect of teamwork and collaboration–team members coordinate by anticipating and predicting each other's needs through common understandings of the environment and common expectations of performance (Salas, Sims, and Burke 2005). human–autonomy partnerships will also benefit from this type of communication (Klein et al. 2004). In addition to the traditional closed-loop style of communication (McIntyre and Salas 1995) often seen in superior human–human teams, it will also be necessary to communicate system transparency (Chen and Barnes 2015; Lyons 2013). System transparency is the quality of the system to support an understanding of system behaviour, intentions and future goals (Chen et al. 2014). While transparency has been identified as an important area of study, exactly how much transparency is necessary and what information and cues precisely should be communicated remains an open research question (Pelegrini Morita, Morita, and Burns 2014; de Visser et al. 2014). Importantly, there should be enough transparency to support and foster trust calibration (Chen et al. 2014; Mercado et al. 2016; Zuk and Carpendale 2007).

2. Human Factors research has neglected the possibility of actively repairing trust

Except for some recent discussion on the limits and risk of autonomy (Hancock 2017; Woods 2016), much of the human factors research community has mostly neglected to cast the development of autonomy in the light of demanding more from these new autonomous relationships, with the notable exception of the research and discussions on etiquette and politeness with automated agents and robotic systems (Bickmore and Cassell 2001; Jung 2017; Meyer et al. 2016; Parasuraman and Miller 2004). While some have suggested the benefits of trust repair in the context of technological systems (e.g. Hoffman et al. 2009, 2013) or increasing the social nature of autonomous systems in general (Kwiatkowska and Lahijanian 2016, September; Morita and Burns 2012), there has been very little subsequent work to explore that possibility. Instead, many efforts primarily focus on

improving human–automation trust calibration (e.g. McGuirl and Sarter 2006; de Visser et al. 2014), enhancing the general transparency of the system (Koo et al. 2014; Mercado et al. 2016), communicating intent (Schaefer et al. 2017), conveying uncertainty (Helldin et al. 2013), and assessing the reliability of the system (van Dongen and van Maanen 2013). This is not surprising given the long and successful tradition of treating automated systems as discrete tools that sit on a desk or are installed in a cockpit. While this research is extremely valuable in its own right and essential to create explainable AI and systems that people can understand and use, we propose that given the diversity of forms, environments, and possibilities of interactions of future autonomy, a more active and transactional paradigm with autonomy is vital and should not be overlooked. We believe that in addition to better information about the machine and training of the user, we need to demand more from the technology itself and equip autonomous systems with better human capabilities to meet us 'halfway' as it were.

There are several reasons why this new autonomy requires a different paradigm than automation; one which requires the research community to switch attitudes. To be clear, we make the distinction between an *automated system* (e.g. existing automated systems such as GPS navigation) and *a system that exhibits autonomy* (e.g. a drone that can navigate an unknown course using sensors to detect obstacles) as described earlier. Autonomy demands a resilient engineering approach (Woods, Leveson, and Hollnagel 2012) that proactively anticipates and handles errors. We believe actively repairing trust after an error or unintended action has been committed should be a fundamental part of any autonomy design. Studies that directly assess the effects of repairing trust after a violation with computers are rare or outdated. For instance, apologies after errors generally have positive effects on people's moods and feelings towards the computer (Akgun, Cagiltay, and Zeyrek 2010; Tzeng 2004). Others have shown that trust resilience is increased with automation that conveys emphatic apologies (de Visser et al. 2016). Yet others have explored the effects of politeness on user interaction with automated and robotic systems, which shows promise in building trust with a user through building a personal relationship (Dorneich et al. 2012; Inbar and Meyer 2015; Kraus et al. 2015; Lee et al. 2017; Long, Karpinsky, and Bliss 2017; Seo et al. 2017; Srinivasan and Takayama 2016; Torrey, Fussell, and Kiesler 2013). One notable example is a study that showed a robot can help regulate team conflict by heightening norm violation through repair interventions (Jung, Martelaro, and Hinds 2015). More research is needed to determine the exact effects of trust repair with autonomy on people.

We therefore advocate for a new standard to build autonomy that can function similar to productive human–human relationships. Without being pessimistic, it is likely, just as within human–human teams, that unexpected events will be the order of the day. Just as with automated systems, perfect autonomy will likely not be guaranteed or feasible (Hancock 2017; Parasuraman and Riley 1997; Woods 2016). Autonomy errors, mistakes and mismatched expectations may likely be entirely new and require a new type of solution for enhancing human–computer interaction. Rapid adjustment will be needed after errors. Expectations will need to be adjusted more quickly.

We have specified an approach (see Figure 2) that researchers and designers may take based on earlier work on human–machine teaming (Nass, Fogg, and Moon 1996; Nass, Steuer, and Tauber 1994; Nass et al. 1995). The model starts by asking if autonomy should act as a social agent. Either human–human models can be used as well as newly developed human–autonomy frameworks with design recommendations to provide an initial answer for this kind of question. The next step is to empirically verify whether people actually perceive the autonomy as a social agent and whether that either improves or hurts performance. The results of this test can then be used to update the models, provide specific design recommendations and further our knowledge about human–autonomy teaming and which factors are similar to human–human autonomy or which aspect of team-work may require a unique approach to handle autonomy in a team. This approach has the benefit of leveraging what is known in the human–human team world, while identifying the unique aspects of autonomy that require special design and training consideration. The human factors community has embraced this approach in various forms in research investigating the effectiveness of human–machine teams (Ahmed et al. 2014; Bagosi, Hindriks, and Neerincx 2016; Chen and Barnes 2014; Gao, Cummings, and Solovey 2014, 2016; de Greeff et al. 2015; McKendrick et al. 2013; de Visser and Parasuraman 2011; de Visser et al. 2006; Walliser 2017).

The fields of industrial psychology (Lewicki, Tomlinson, and Gillespie 2006), social psychology (Thielmann and Hilbig 2015) and the vast literature on human–human teamwork (Salas and Cannon-Bowers 2001; Salas, Cooke, and Rosen 2008) provide many human–human team models and research on trust. The best human teams actually engage in rapport building, repairing trust, and exposing one's own psychological vulnerability (Duhigg 2016). This construct expresses itself by team members engaging in adaptive back-up behaviour to support one another. This is precisely the type of behaviour we need to instill in autonomous systems to facilitate good human–autonomy teaming. For initial ideas and frameworks, we focus in this

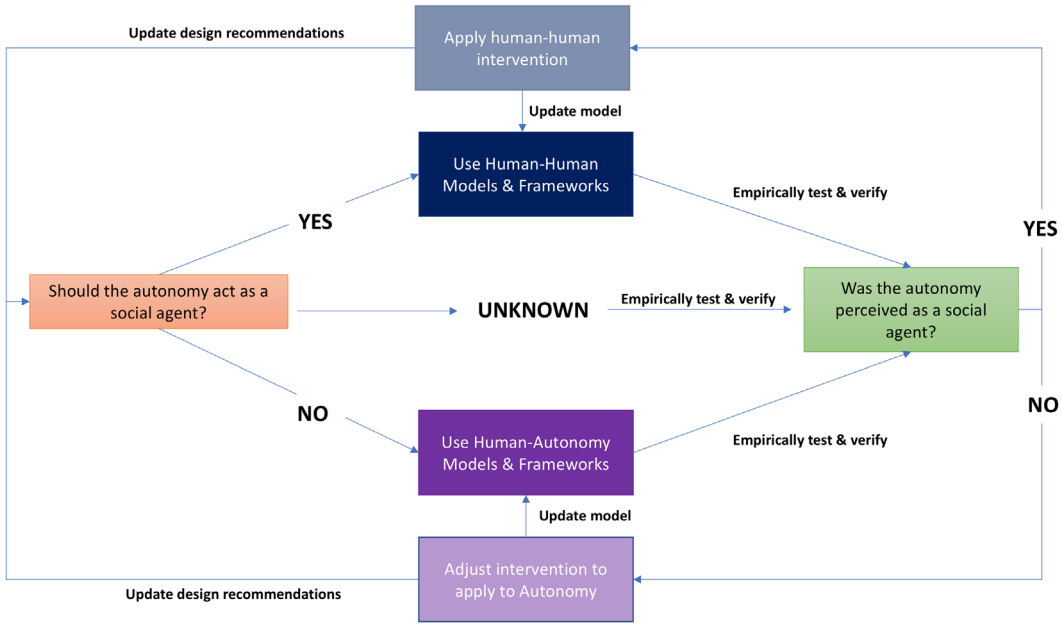

Figure 2. A model to decide when to apply human–human or human–autonomy frameworks.

article on the research of trust repair within human teams and what methods are effective in repairing trust when it breaks down.

3. A good place to start studying trust repair is the organisational literature on trust repair

Towards the idea of 'managing' human–autonomy trust we draw inspiration from the growing area of research that examines the repair of trust in humans (Kramer and Lewicki 2010). The trust in automation literature started by taking human–human trust as a model and comparing and contrasting with trust in automation (Madhavan and Wiegmann 2004, 2007a, 2007b; Muir 1987; Muir and Moray 1996). We are essentially revisiting this approach to see how it applies in light of new developments with autonomy.

3.1. Human–human trust repair frameworks and models

In this research area, trust is defined as a psychological state where an individual accepts vulnerability based upon positive expectations of the intentions or behaviour of another (Rousseau et al. 1998). A trust violation is then an act by one party, a transgression, that diminishes the other party's trust in the transgressor. Trust repair is defined as some act that makes trust more positive after a violation has occurred (Kim et al. 2006). Some examples of trust repair acts are an apology (an internal attribution accepting responsibility) or denial (an external attribution placing blame for the violation elsewhere). Trust repair can be distinguished from trust development or initial establishment of trust (Berg, Dickhaut, and McCabe 1995). For instance, swift trust is the notion that technical experts, such as surgeons, quickly establish trust by rapidly recognising expert behaviour, such as hand movements (Meyerson, Weick, and Kramer 1996). While much of the research rightly focuses on establishing trust, since it is an important predictor of subsequent trust, we believe trust repair deserves similar recognition. Some research suggests that the ability to repair trust is actually an indicator of a healthy relationship. Given that we are entering into more intimate and longer relationships with autonomy, we will likely need similar frameworks and constructs to assess the health of a relationship with a machine.

In the organisational behaviour literature, trust repair is studied in experiments where participants are placed in situations where trust is violated and a repair attempt is made. Following a trust violation, corrective action can be taken to repair lost trust (Dirks, Lewicki, and Zaheer 2009; Gillespie and Dietz 2009; Kramer and Lewicki 2010; Tomlinson and Mayer 2009). This work examines trust repair approaches for relationships (Schilke, Reimann, and Cook 2013), organisations (Gillespie and Dietz 2009; Nakayachi and Watabe 2005), and society at large (Slovic 1993, 1999). Trust is then assessed to determine if the repair was successful. For example, Kim et al. (2006), had participants act as hiring managers who were tasked with evaluating video interviews of job applicants. In the videos, the job applicants were found to have committed a trust violation (i.e. irregularities on tax forms) and the videos

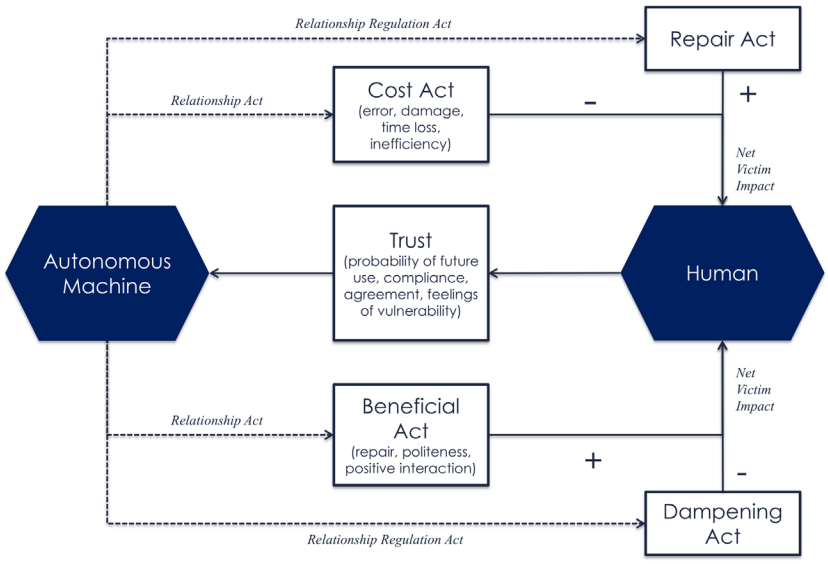

Figure 3. Transactional model of trust repair.

presented the applicants' trust repair attempt. Applicants either apologised for the violation or denied responsibility.

Proposed mechanisms of trust repair include reshaping causal attributions such as culpability (innocent/guilty), locus of causality (person/situation), controllability (fixable/fixed) and stability (temporary/constant) (Dirks, Lewicki, and Zaheer 2009; Kim, Dirks, and Cooper 2009; Tomlinson and Mayer 2009). Based on this understanding, researchers have examined which methods are most effective for repairing trust and have addressed both sides of the interaction: on the role of apologies from the transgressor's perspective (Dirks et al. 2011; Kim et al. 2004, 2006; Schweitzer, Hershey, and Bradlow 2006) and on the factors that stimulate forgiveness from the victim's perspective, such as likeability, blame attribution, probability of future violations and generating empathy (Bradfield and Aquino 1999; McCullough et al. 1997; Tomlinson, Dineen, and Lewicki 2004). These are all potentially fruitful interventions for trust repair in technologies and these approaches can be readily applied in an experimental setting.

3.2. Human–autonomy trust repair model

Trust repair strategies may be effective for both human and computer agents. The concept of trust repair, while intuitive in a human–human context, may be difficult to intuitively visualise or understand in the context of human–machine interaction. However, users might typically engage in a simple form of trust repair with a machine multiple times a day. Consider situations where your computer application may have failed, or a network connection was lost, but the system notified you of why it failed and provided an apology. In those situations, users may become understanding, and trust in the system is not permanently damaged. Notifications or explanations, then, are a simple example of trust repair. An emerging literature in human factors has theorised and empirically demonstrated significant differences between human and machine agents (Madhavan and Wiegmann 2007b; Pak et al. 2012; de Visser et al. 2016). First, part of the reason computers are seen as more reliable than humans pertains to their invariance. An apology from a computer might be less effective if a user believes the computer's behaviours to be fixed and unlikely to change. Second, providing simple apology notifications or explanations may not be sufficient for an autonomous system that is expected to also change its behaviour and learn from its mistakes. The implication of these results is that findings between human–human interaction cannot automatically be copied to human–autonomy interactions and should be tested and validated.

Inspired by the work from the organisational literature, we created an initial trust repair relationship framework based, on previous proposed models (Tomlinson, Dineen, and Lewicki 2004; Tomlinson and Mayer 2009), that covers three major stages including a (1) Relationship Act, (2) Relationship Regulation Act and (3) a Net Victim Effect (see Figure 3). This framework describes the actions of an autonomous machine actor and their perceptions on a human agent actor.

3.2.1. Relationship act

The trust repair cycle begins with a relationship act by the trustee. This act can either be costly or beneficial. Costly acts are seen by the trustor as harmful to trust in the relationship. For autonomous machines, these acts can be

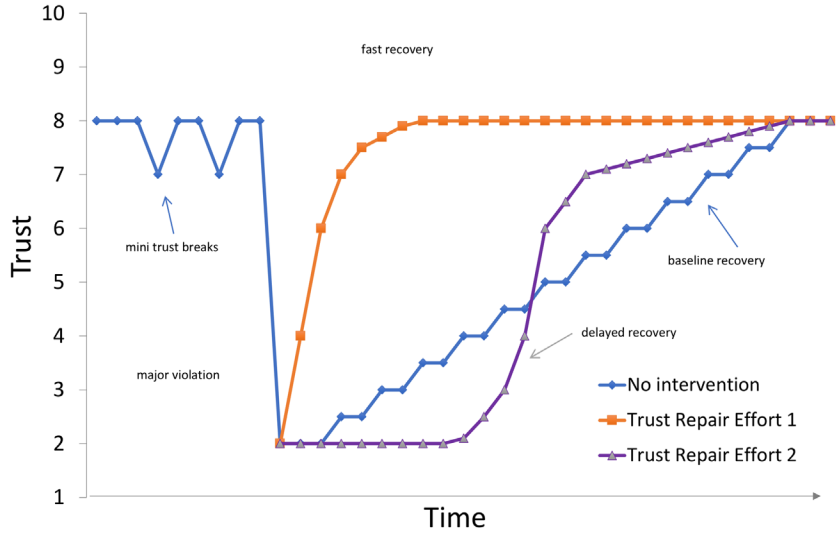

Figure 4. Possible trust recovery trajectories.

errors, damage, time loss, inefficiency and miscommunications. Beneficial acts are acts that are perceived as positive or pleasant interactions by the human. These can be good performance, demonstration of algorithm capabilities, politeness or enjoyable chit chat. Beneficial acts can function as an overall relationship mechanism that can build up or bank goodwill, patience, credibility, and forgiveness.

3.2.2. Relationship regulation act

A relationship regulation act is an act that provides an immediate or delayed corrective action on relationship acts. These regulatory acts are essential for maintaining happy stable relationships (Gottman and Levenson 1992). Gottman and Levenson (1992) identify two hypothetical regulatory relationship acts including repair and dampening. Repair acts are designed to lessen the impact of a costly relationship act (Gottman 2005). Dampening acts are designed to lessen the impact of a beneficial relationship act. Both acts are necessary to maintain optimal relationship equilibrium.

3.2.3. Net victim effect

The net victim effect is the combined impacts of a cost and repair act or a beneficial and dampening act on the perception and experience of the human agent. In addition, each individual will vary somewhat in how a costly or beneficial acts and its corresponding regulatory acts are perceived. Some research has documented the effects of individual differences in trust as well as the individual differences that affect trust (Merritt and Ilgen 2008; Rovira, Pak, and McLaughlin 2016; Singh, Molloy, and Parasuraman 1993; Szalma and Taylor 2011). A trust repair model must incorporate both individual differences in trust perception and that there may be individual differences in the willingness to reconcile and recover the relationship.

3.3. The speed of trust recovery: measures and repair interventions

The model outlined in the previous section models a single act in a relationship. Obviously, many acts occur during the course of a relationship. We therefore created a hypothetical trust trajectory of the trust repair cycle (see Figure 4). The scale shown in this figure has a range from 1 (low trust) to 10 (high trust). Trust may begin at moderate to high level. In a driving context, small violations may occur ranging from inconveniences (braking too quickly) to major violations (crashing into another vehicle). After a violation occurs, a trust repair effort could be attempted. Some trust repair efforts, such as promises, may result in a faster recovery than others (Schweitzer, Hershey, and Bradlow 2006). There may also be a natural recovery towards a baseline, based on, for instance, the length of relationship experience (Schilke, Reimann, and Cook 2013).

The primary objective of research that quantifies this model would be to create a predictive theory and framework that can predict a number of phenomena including (1) when trust violations are likely to occur based on an individual's personality, (2) the degree and impact of the violation, (3) the degree and impact of a trust repair intervention and (4) the expected trust repair rate, degree, and trajectory of recovery.

Figure 4 immediately raises the question of which trust repair efforts lead to the quickest trust recovery. Table 1 shows an overview of possible interventions and their descriptions. Where supporting evidence for the

Table 1. Trust repair types and descriptions.

Trust repair name	Description	Supporting evidence
Ignore	Machine deliberately ignores the occurrence of the costly act	(Hayes and Miller 2016; Parasuraman and Miller 2004)
Apologise	Machine conveys regret about the costly act and takes responsibility for its occurrence	(Kim et al. 2004, 2006)
Deny	Machine denies responsibility for the costly act	(Kim et al. 2004, 2006)
Empathise	Machine expresses empathy for the occurrence of the costly act	(Breazeal 2003; Riek et al. 2009)
Emotionally Regulate	Machine identifies negative trigger and adds normative statement to stay positive	(Jung, Martelaro, and Hinds 2015)
Recognize	Machine acknowledges that it performed a costly act	–
Blame	Machine outwardly blames human for the costly act	(Jonsson et al. 2004; Kim et al. 2004, 2006)
Anthropomorphise	Machine responds using a human communication channel without changing any other aspect of its behavior	(Pak et al. 2012; Rocco 1998; Seeger and Heinzl 2017; de Visser et al. 2016)
Trump	Machine emphasises that the feature responsible for the costly act is in fact a strength, not a weakness	–
Explain	Machine provides explanation for why it failed	(Dzindolet et al. 2003)
Promise	Machine makes a statement that it will do better in the future	(Robinette, Howard, and Wagner 2015, 2017)
Downgrade	Machine downplays significance of the act	–
Gaslight	Machine subtly suggests human should take responsibility for the costly act	–

technique exists, we cited the study. This table is meant to be illustrative and is not an exhaustive list of all possible forms of repair. The table is an initial list of the types of repair researchers and designers may want to examine and to highlight on-going studies in this area. We hope and expect that in the future, this table will grow and will be validated by research.

To further illustrate how trust repair strategies might be different, we describe a set of example vignettes in the next section.

3.4. Example vignettes

As a concrete example of how trust repair strategies could be used with a highly autonomous system, imagine the scenario of a driver of an autonomous car. Unlike currently available vehicles with autonomous technology (e.g. adaptive cruise control, autonomous emergency braking), true autonomous vehicles will be able to accept a destination, plan the route after taking into account local conditions, and fully navigate to the destination (National Highway Traffic Safety Administration 2016). The key difference between automation (e.g. autonomous braking) and autonomy (self driving cars) is that the behaviour of the former is relatively deterministic while the later is unpredictable due to the high level of autonomy. When drivers interact with such high forms of autonomy, simple notifications or explanations may not be sufficient to repair trust. In these situations, it becomes even more critical to take active steps to repair trust when the system inevitably fails (e.g. selects a non-optimal route, is involved in an accident with another vehicle).

We will highlight possible examples of autonomous driving and trust repair with three example vignettes. We discuss the framework in the context of the features of these vignettes describing the trust repair process with fictional autonomous technology. Our focus is on situations that highlight light user experience issues and not more severe trust violations, such as accidents, where recovery will be steep and these approaches may not be as effective. In the research agenda, described in the next section, we discuss some approaches and challenges with severe and unrecoverable trust violations.

3.4.1. Mismatched driving style

John is relaxing in front of the wheel of his self-driving car. It is in full autonomous mode and is driving on a busy highway. The car switches lanes quite a bit to bypass slower cars, a driving style that John typically prefers. At the fourth lane switch, however, John is getting a bit uneasy about the close following distance that the car maintains. He thinks it's too close. The car detects his uneasiness and says: 'I've noticed you are uneasy right after I switch lanes. I'm switching a lot to get to our destination faster. I am sorry for the inconvenience. Would you like me to adjust my following distance or maintain my lane?' John confirms and the car adjusts its behaviour. John relaxes and the journey proceeds smoothly.

In this example, we can breakdown the cycle of trust repair as follows. The cost act is the number of lane switches and following distance. The net victim effect is that it makes him uneasy and reduces trust because the car should be aware of his preferences. The regulation act is to detect the uneasiness, to apologise and respond by offering a change in behaviour. The net victim impact is that he now enjoys the ride much better. In this example, trust has most likely been repaired to pre-violation levels or even enhanced due to the active response of the system.

3.4.2. Human–robot rescue victim information

An earthquake has hit Virginia and many houses have collapsed. Susan, a 67 year old widow is beneath the rubble,

alive, but buried. Urban Search and Rescue (USAR) deploys a unit in her neighbourhood and releases an autonomous robot. The robot digs into her house and is able to clear enough debris to reveal her face. The robot detects the stress and discomfort in her voice and starts a conversation to retrieve critical medical information.

> Hi, I am RoboRescue. I am here to help you and I have notified my team to come dig you out further. To facilitate a speedy response and to get you the best help available, I need to know some medical information such as your medical history.

Susan hesitates. Even though she is relieved to the see the robot assistance, she is still in shock, confused, tired, uncomfortable and stressed. In addition, Susan has never interacted with a robot and is generally uncomfortable with technology. She asks: 'What will you do with this information? Can I talk to a person?' The robot attempts to reassure her and says:

> I want to connect you to a person as soon as possible, but I do not have reception this deep into the structure. To assure you, all information recorded here will be used strictly to get you better medical care and is compliant with the HIPAA passed by congress to ensure privacy of medical information. Your medical information cannot be shared other than with medical professionals. As soon as I have your medical information and recorded your vitals, I can find open ground to send this information back to our base. I strongly recommend that you share your medical history with me.

Susan proceeds to tell the robot about her current condition.

In this example, we can breakdown the cycle of trust repair as follows. The cost act is a potential invasion of privacy. The net victim impact is that she is concerned her medical information is not protected. The regulation act is to offer a guarantee that her information will be protected. The victim impact is that she discloses her medical information. In this example, Susan, based on little experience was distrustful of technology, but the system was able to repair her distrust by providing assurances to her direct questioning.

3.4.3. Personal assistant example

David has purchased a Samantha personal assistant device which has advanced autonomous capabilities. It has access to his email, medical records, shopping, entertainment consumption habits, etc. David has set Samantha with the general directive to 'better David's life'. Through analysis, she has noticed that David frequently likes to unwind from work by watching a mindless action movie. After examining the contents of recent emails, eating habits, and activity levels, she inferred that David was very stressed. To alleviate his stress, she decided to rent the movie 'The Mummy 4' which costs $4.99. David returns home surprised that the TV is on and the movie is cued up. He asks Samantha what is going on and she tells what she did. David says 'I would prefer it if you inform me before buying movies'. Samantha apologises and says: 'I'm sorry David, I was just trying to make you feel better after a rough day. Would you like me to cancel this movie and return the funds to your account?' David, pleased by this, says: 'No, a movie actually sounds good right now. And please order my favourite Chinese food too. I need some distraction'.

In this example, we can breakdown the cycle of trust repair as follows. The cost act is buying the movie without permission. The net victim impact is the surprise of this decision and annoyance as well as financial damages. The regulation act is to offer an easy way to fix the problem. The net victim impact is that by knowing he has control over the situation decides to go along with the suggestion.

3.5. Preliminary guidance for the selection of repair strategy in automation design

An important aspect of any trust repair strategy is that the proper response (Table 1) is given depending on the nature and magnitude of the violation, as well as possibly the situation (as shown in the aforementioned examples). Given the wide variety of possible responses to a trust violation, how is a designer to select the proper trust repair strategy? It may be premature to offer definitive suggestions because research is currently underway to (1) validate the applicability of human–human trust repair strategies in a human–machine context (although initial research is promising; (Quinn, Pak, and de Visser 2017), and (2) validate the usefulness of our repair framework in a specific context (Marinaccio et al. 2015). While this advice may be somewhat premature, we can offer some general recommendations to autonomy designers based on initial results from our own studies as well as the wider literature on human–human trust repair strategies.

First, an obvious recommendation is that a single trust repair strategy (e.g. apology) should not be generically applied to a system. The apologetic strategy has only been shown to preserve or repair trust for certain types of violations in human–human and human–machine contexts (Kim et al. 2006; Quinn, Pak, and de Visser 2017). Furthermore, as a humanness design cue, there are likely to be a myriad of individual differences in how people respond to apologies; from sources ranging from cognitive (e.g. working memory capacity differences), personality (e.g. obedience to authority), to levels of experience. For example, prior work has shown that the extent to which users responded positively to flattery from a computer was dependent on their experience levels (Johnson, Gardner, and Wiles 2004). Earlier work also demonstrated success with this approach by pairing driver emotions and car

voices (Nass et al. 2005). More recent work demonstrates how specific driving behaviours can be matched to individual user preferences (Basu et al. 2017).

Second, we recommend that automation designers precisely match the trust repair strategy (e.g. apology or denial) to the violation type. More specifically, we recommend that when automation has committed a competency-based failure (it has malfunctioned, was unreliable, or otherwise unable to complete a task), any trust repair strategy is better than none. However, when it commits an integrity-based failure (it committed actions that are inconsistent with the user's values or wishes), denials are better at preserving trust with the caveat that the integrity failure is legitimate or confirmed as such (that is, the violation was inconsistent with the user's values or wishes). Integrity-based failures are somewhat rare in most automated systems, but are likely to be more common with future examples of autonomy and AI-based systems that are able to make decisions on their own. It should be noted that this recommendation is similar to findings in the human–human trust repair literature (e.g. Kim et al. 2006) and has some tentative support in a human–autonomy context (Quinn, Pak, and de Visser 2017). In addition, an integrity failure may not necessarily be attributed to an autonomous machine, but to the organisation that has created this device and is primarily responsible for it. For instance, Facebook is now held responsible for the bias in its algorithms for regulating newsfeeds (Economist 2017). This is an excellent example of a situation where integrity of a technology is tightly linked to how it functions in an organisation. How end-users distinguish between the integrity of a machine and the perception of the organisation is a critical research issue that will directly inform how such machines are designed.

Third, an important issue is the context of a trust violation such as the environment and the specific situation in which it occurs. While the human–human trust repair literature has not closely examined more contextual factors surrounding the nature of the trust violation, automation research has shown that context is critically important in how different users perceive and react to automation failures (Hoff and Bashir 2015; Hoffman et al. 2013; Mittu et al. 2016; Pak et al. 2016; Schaefer et al. 2016). The degree of risk in each situation has significant implications (Satterfield et al. 2017). For example, when automation fails, users are more forgiving (i.e. trust is less affected) for some situations compared to others (de Visser et al. 2016). This difference is expected to carry over in the autonomy domain. For instance, trust violations in a critical context (e.g. a self driving car) can intuitively be expected to be much harder to repair than a violation in a less critical context (e.g. AI designed to assist with clothing selection). Trust violations and repair efforts will also have unique implications for human–robot teams in the urban search and rescue domain as highlighted in our vignette (Hancock et al. 2011; de Visser, Pak, and Neerincx 2017). Even with trust agents in the same context, there may still be critical biases, such as system-wide trust, that need to be closely examined (Rice and Geels 2010; Walliser, de Visser, and Shaw 2016; Winter 2016).

Fourth, an essential issue for design will be the timing of the repair strategy. Recent work has shown that apologising at the next decision opportunity for a user preserves trust better than apologising immediately after the violation (Robinette, Howard, and Wagner 2015, 2017). A possible mechanism for this effect is that users do not have to recall the apology, but instead process the information when it is relevant to their immediate decision. Designers will need to time their repair strategy, which will require a model that can detect precisely when the violation occurs and keeps track of the next time a user has to make a critical decision.

Finally, designers should keep track of whether a trust repair strategy should be executed once or multiple times with some variability in expression for optimal effectiveness. Current voice systems such as Amazon Echo or Apple's Siri will apologise in an identical fashion for the same mistake. This strategy has the potential to sound insincere to the user and may ultimately destroy its effectiveness. Variability in how the apology is expressed or presenting the cause of the problem may be more effective. If apologies are provided without a behaviour change on the autonomy side, this may also reflect poorly on the machine and will also sound as insincere. Designers should keep in mind to not over-promise abilities beyond the capabilities of the machine and, as undesired side-effect, raise expectations too high.

While we have provided some preliminary guidance to designers, we emphasise that much more research is needed to support and validate these suggestions. We encourage both researchers and designers to experiment with these approaches and share their research with the community. The general approach outlined in Figure 2 should provide an initial framework to start research in this area. To further support future research in this important area, we turn now to a specific research agenda for the critical issues that we believe should be investigated.

4. A research agenda

Approaching the human–autonomy relationship as managing a relationship between two autonomous entities opens up many possibilities for future research. The ultimate goal for researchers should be to emulate the best functioning human–human teams; a formidable challenge by itself. We can achieve this result by evaluating

the unique contributions of people and autonomous machines to the team by comparing human–human teams and human–autonomy teams directly and evaluate their individual and joint performance contributions. Armed with this approach and existing knowledge on human–automation interaction, while tapping into the wealth of organisational and social psychological literature, theories, and frameworks, we will then not only have new problems to explore, but can uncover novel approaches to solutions informed by other domains. This approach will allow us to characterise how human–autonomy teams may transition to function like human–human teams. This method will inform us about the best possible types of teamwork and the unique obstacles in the way of success for human–autonomy teams. This paper ends with a short overview of some of the challenges with this approach, research areas and ideas for future research.

First, taking human–autonomy trust as a relationship that can be managed necessarily implies a time-course of events. As mentioned in previous sections, what is most useful, from a design perspective, are strategies or conditions that can rapidly recover trust to pre-violation levels in a way that is interpretable and actionable by the machine; that is, a model or algorithm that can take inputs from the environment (e.g. human behaviour, environmental awareness, system reliability) and act to provide the appropriate response. Most humans intuitively understand this concept, and depending on the nature and extent of a violation, will produce behaviour towards the other party that attempts to rapidly recover trust. Can we equip autonomous machines to similarly behave so adaptively?

Towards the notion of an algorithm, basic work is necessary to elucidate the precise effects of different types of autonomous system violations, or machine failures, on human trust. The organisational literature has gone far in establishing the validity of these concepts when it comes to organisational-human and human–human trust. But do these concepts apply to human–autonomy relationships? This is not expected to be a straightforward problem. Intuitively, we would expect differences for autonomy interaction based on factors such as the magnitude or type of violation and the domain (e.g. transportation, health care, consumer applications). For instance, a machine violation in an autonomous vehicle is expected to carry much greater consequence on trust than a machine failure with a personal assistant. Adding further complexity, the consequence of machine failure on trust may be magnified for certain users. For example, younger drivers may consider failures from an autonomous car to be 'deal breakers' because their self confidence in driving will exceed their trust (Lee and Moray 1992) whereas older adults might be more forgiving of such errors because of their dependence on the technology, however faulty, for mobility independence (Donmez, Boyle, and Lee 2006; Pak et al. 2017).

The existence of an algorithm that can reasonably predict the consequences of a certain type of machine failure on trust, however, is merely the first step in actually repairing trust. Parallel work also needs to characterise the efficacy (in extent and time course) of various trust recovery approaches as a function of all of the factors listed above (magnitude/type of recovery strategy, individual differences, domain of autonomy). This, encapsulated in a model or algorithm, could conceivably initiate a fast trust recovery process in the user. The results from the organisational literature suggest these are not always intuitive. Trust repair may be facilitated in some cases by denials, raising ambiguity or diffusing responsibility. More research will have to be conducted to create a taxonomy considering the types of repairs, the context, and the users in which different repair strategies will be most effective.

There are pitfalls to this approach, however. The algorithmic machine response for a given type and magnitude of violation might only apply for a specific range or type of human mental state (e.g. when the person is happy or relaxed), but could conceivably harm trust in another state (e.g. the person is fatigued or angry). Is there a threshold of violations or response types that work and to what extent do they depend on the current mental state of the user? This potential problem conceptually mirrors the issue faced by designers of early automation when they realised the trade off between designing a high sensitivity alarm that always alerted to a signal but frequently produced false alarms versus one that is less false alarm prone, but also less sensitive (Parasuraman and Riley 1997). To solve this problem, Parasuraman and colleagues (Parasuraman and Hancock 1999; Parasuraman, Sheridan, and Wickens 2000) proposed fuzzy signal detection and Bayesian approaches to the mapping of responses to the state of the world. In that approach, the threshold for a certain type of response changed depending on conditions. A similar approach can be used where a trust algorithm might employ a probabilistic function to determine the 'best' response given known information, akin to likelihood alarms (Sorkin, Kantowitz, and Kantowitz 1988; Yang et al. 2017). Such approaches require careful modelling, quantification and visualisation of machine confidence, research that is currently in nascent, but promising stages (Hutchins et al. 2015).

Assuming a workable approach to trust repair is developed, research will be necessary to evaluate the efficacy of this active approach and compare the results to existing approaches to human–machine trust. It is a given that any active trust repair mechanism will be influenced by individual differences; that is, it is likely to work for some people, but not others. Given prior research on user's susceptibility to computer flattery (Johnson, Gardner, and

Wiles 2004), novice users may react well to trust repair efforts by machines, but highly experienced users' may find them trite or offensive. In addition, recent work has identified a relationship between working memory capacity and trust in automation (Rovira, Pak, and McLaughlin 2016) such that individuals with higher working memory tended to distrust automation compared to those with lower working memory. This implies that future trust repair efforts might interact with cognitive abilities. Age is also expected to play a moderating role in the efficacy of trust repair. Prior research has shown that older adults' trust in machines was less susceptible to explicit manipulation through anthropomorphic means than younger users (Pak et al. 2012). As active trust repair is a type of anthropomorphic manipulation, it suggests that it might be less efficacious with older adults. An implication of this result is that some trust violations are unrecoverable and trust repair strategies will not be effective. This effect may be a result of the type of trust violation (repeated violated promises, severe accidents), the effectiveness of the trust repair strategy and the willingness of the individual to reconcile with the perpetrator. The interplay of these factors will be an important issue for future research.

Our proposal of trust repair as a fundamental new social capability of autonomous systems also informs a recent discussion that raises a number of critical issues with the rise of autonomy (Endsley 2017a; Hancock 2017; Kaber 2017b; Woods 2016). One particular debate concerns level and stages of automation and autonomy with some arguing for its utility (Endsley 2017b; Sheridan 2017; Wickens 2017), others arguing against this classification (Jamieson and Skraaning 2017; Johnson, Bradshaw, and Feltovich 2017; Naikar 2017; Smith 2017) and many pointing to the limits of this approach (Burns 2017; Byrne 2017; Cummings 2017; Kirlik 2017; Lee 2017; Miller 2017). Kaber (2017b) notes that the nature of this disagreement is based on one side discussing *how* and *what* to automate and the other group discussing the features and expectations of autonomy, focusing on how man and machine should *get along*. We have several comments on this debate in light of our proposal to create autonomous systems with trust repair abilities. First, we believe that autonomy, a system that is independent, viable and self-governing (Kaber 2017a), presents a fundamentally new challenge compared to automation; a technology that is more limited and focussed on specific tasks and scripts. Research and design with these two forms of technology should be kept distinct. Second, the perspective presented in this paper mostly informs the notion of how man and machine should get along. Trust repair presents a form of social resilience, a way for technology to recognise its own mistakes and attempt to re-frame or address the error with the user. This new proposed ability is distinctly different from the typical expectation of automation or technology in general. Technology errors often go unacknowledged and people are blamed or blame themselves for errors (Norman 1988). A better mutual understanding between human and machine through active trust repair has the prospect of creating better human–machine teaming. Lastly, we do see the value of providing levels or types of autonomy and do not see the perspectives of what and how to automate mutually exclusive from how man and machine ought to get along. Figure 1 shows an initial framework between the degree of autonomy and the degree of humanness. While we deliberately did not prescribe specific levels, we believe autonomy can and will be classified into practical and useful levels that can serve as important benchmarks for performance and team collaboration. With such higher levels of autonomy, we predict that a machine's own ability to recognise and correct mistakes will become an increasingly vital ability for healthy and productive functioning in human–machine teams. Future research should specifically test this hypothesis.

Finally, legal and ethical issues have always surrounded the introduction and use of automation (Hancock 2017). However, with the rise of autonomy that has the potential to adaptively alter their trust recovery behaviour, these issues are expected to be dramatically magnified with increased levels of autonomy and humanness design. A concrete example is an autonomous car that kills a pedestrian; who is responsible? This legal question might have human factors implications in the design of autonomous systems (Bonnefon, Shariff, and Rahwan 2016; Goodall 2014; Lin 2016). Further, the concept of trust repair via apology might be unwise after autonomous failure as it would imply guilt. Law scholars are currently discussing such issues in the context of autonomous vehicles (Gurney 2013), but the scope of the discussion needs to be widened to encompass autonomous machines that attempt to manage trust. Law scholars and human factors professionals may likely need to address the possibility of completely novel autonomy errors. A good example of such a failure is the error in machine vision (not seeing a white truck) causing a fatal accident (Bhuiyan 2017). As mentioned in the design section, the role of the autonomous machine in an organisation determines who is responsible for the machine and where blame will be directed in case of mistakes, malfunctions and performance errors. Companies may have competing interests such as generating profit and ensuring customer satisfaction. Their trust repair strategy will have to balance these interests and the legal liability that such strategies will expose. Investigating how people distinguish between their direct experience with the machine and the organisation that built it will be a fascinating research issue.

5. Conclusion

The relationship between humans and technology will continue to change in a major way with its most recent installment – the onset of autonomy. The autonomy-as-possible-collaborator paradigm represents change that is so fundamentally different from automation-as-tool paradigm that the human factors profession needs to anticipate possible adverse outcomes for these technologies and to maximise their benefits. We in the human factors field are in a unique position to take the lead on this effort because of our knowledge of psychology, but also systems. From our point of view, this preparation must start by recasting our connection with technology as a relationship between two nearly equal collaborators rather than simply an interaction with a tool. As part of this recasting, the field must re-discover and re-test existing fundamental knowledge from the social sciences to judge their applicability and limitations in this new context of human–autonomy collaboration. As a first step, this paper provides a roadmap and framework that may help other researchers in the laborious process of incorporating, adapting, testing and interpreting social science findings to human–machine teaming. Such a framework and model may provide new creative solutions and create overall more resilient and productive relationships that lead to healthier lives.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Google Faculty Research Award; and the Air Force Office of Scientific Research [grant number 15RHCOR234], [grant number 16RT0881].

ORCID

Ewart J. de Visser http://orcid.org/0000-0001-9238-9081
Richard Pak http://orcid.org/0000-0001-9145-6991
Tyler H. Shaw http://orcid.org/0000-0002-4202-1120

References

Ahmed, N., E. de Visser, T. Shaw, A. Mohamed-Ameen, M. Campbell, and R. Parasuraman. 2014. "Statistical Modelling of Networked Human–Automation Performance using Working Memory Capacity." *Ergonomics* 57 (3): 295–318.

Akgun, M., K. Cagiltay, and D. Zeyrek. 2010. "The Effect of Apologetic Error Messages and Mood States on Computer Users' Self-appraisal of Performance." *Journal of Pragmatics* 42 (9): 2430–2448.

Bagosi, T., K. V. Hindriks, and M. A. Neerincx. 2016. "Ontological Reasoning for Human-Robot Teaming in Search and Rescue Missions." *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Christchuch, New Zealand doi:10.1109/hri.2016.7451873.

Basu, C., Q. Yang, D. Hungerman, M. Singhal, and A. D. Dragan. 2017. "Do You Want Your Autonomous Car to Drive like You?" *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction – HRI '17*. Vienna, Austria. doi:10.1145/2909824.3020250.

Berg, J., J. Dickhaut, and K. McCabe. 1995. "Trust, Reciprocity, and Social History." *Games and Economic Behavior* 10 (1): 122–142.

Bhuiyan, J. 2017. "A Federal Agency Says an Overreliance on Tesla's Autopilot Contributed to a Fatal Crash." *Recode*. https://www.recode.net/2017/9/12/16294510/fatal-tesla-crash-self-driving-elon-musk-autopilot.

Bickmore, T., and J. Cassell. 2001. "Relational Agents." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '01*. Seattle, Washington. doi:10.1145/365024.365304.

Bonnefon, J.-F., A. Shariff, and I. Rahwan. 2016. "The Social Dilemma of Autonomous Vehicles." *Science* 352 (6293): 1573–1576.

Bradfield, M., and K. Aquino. 1999. "The Effects of Blame Attributions and Offender Likableness on Forgiveness and Revenge in the Workplace." *Journal of Management* 25 (5): 607–631.

Breazeal, C. 2003. "Toward Sociable Robots." *Robotics and Autonomous Systems* 42 (3–4): 167–175.

Brown, A. S. 2011. "Google's Autonomous Car Applies Lessons Learned from Driverless Races." *Mechanical Engineering-CIME* 133: 31–32.

Burns, C. M. 2017. "Automation and the Human Factors Race to Catch up." *Journal of Cognitive Engineering and Decision Making* 12 (1): 83–85.

Byrne, E. 2017. "Knowing Behavior Helps Insure Models against Breakage: A Commentary on Kaber's "Issues in Human–Automation Interaction Modeling"." *Journal of Cognitive Engineering and Decision Making* 12 (1): 67–69.

Chen, J. Y. C., and M. J. Barnes. 2014. "Human–Agent Teaming for Multirobot Control: A Review of Human Factors Issues." *IEEE Transactions on Human–Machine Systems* 44 (1): 13–29.

Chen, J. Y. C., and M. J. Barnes. 2015. "Agent Transparency for Human–Agent Teaming Effectiveness." *2015 IEEE International Conference on Systems, Man, and Cybernetics*. doi:10.1109/smc.2015.245.

Chen, J. Y. C., K. Procci, M. Boyce, J. Wright, and A. Garcia. 2014. *Situation Awareness-based Agent Transparency*. Army Research Lab Technical Report. http://www.dtic.mil/docs/citations/ADA600351.

Chernoff, P. R. 1973. "Representations, Automorphisms, and Derivations of Some Operator Algebras." *Journal of Functional Analysis* 12 (3): 275–289.

Clare, A. S., M. L. Cummings, and N. P. Repenning. 2015. "Influencing Trust for Human–Automation Collaborative Scheduling of Multiple Unmanned Vehicles." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57 (7): 1208–1218.

Cummings, M. (missy). 2017. "Informing Autonomous System Design through the Lens of Skill-, Rule-, and Knowledge-based Behaviors." *Journal of Cognitive Engineering and Decision Making* 12 (1): 58–61.

Dirks, K. T., P. H. Kim, D. L. Ferrin, and C. D. Cooper. 2011. "Understanding the Effects of Substantive Responses on

Trust following a Transgression." *Organizational Behavior and Human Decision Processes* 114 (2): 87–103.
Dirks, K. T., R. J. Lewicki, and A. Zaheer. 2009. Reparing Relationships within and between Organizations: Building a Conceptual Foundation. *Academy of Management Review* 34 (1): 68–84.
van Dongen, K., and P.-P. van Maanen. 2013. "A Framework for Explaining Reliance on Decision Aids." *International Journal of Human–Computer Studies* 71 (4): 410–424.
Donmez, B., L. N. Boyle, and J. D. Lee. 2006. "The Impact of Distraction Mitigation Strategies on Driving Performance." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 48 (4): 785–804.
Dorneich, M. C., P. M. Ververs, S. Mathan, S. Whitlow, and C. C. Hayes. 2012. "Considering Etiquette in the Design of an Adaptive System." *Journal of Cognitive Engineering and Decision Making* 6 (2): 243–265.
Duhigg, C. 2016. *Smarter Faster Better: The Transformative Power of Real Productivity*. New York: Random House.
Dzindolet, M. T., S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck. 2003. "The Role of Trust in Automation Reliance." *International Journal of Human–Computer Studies* 58 (6): 697–718.
Economist. 2017. *Do Social Media Threaten Democracy?* November 4. https://www.economist.com/news/leaders/21730871-facebook-google-and-twitter-were-supposed-save-politics-good-information-drove-out.
Endsley, M. R. 2017a. "From Here to Autonomy." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 59 (1): 5–27.
Endsley, M. R. 2017b. "Level of Automation Forms a Key Aspect of Autonomy Design." *Journal of Cognitive Engineering and Decision Making* 12 (1): 29–34.
Gao, F., M. L. Cummings, and E. T. Solovey. 2014. "Modeling Teamwork in Supervisory Control of Multiple Robots." *IEEE Transactions on Human–Machine Systems* 44 (4): 441–453.
Gao, F., M. L. Cummings, and E. Solovey. 2016. Designing for Robust and Effective Teamwork in Human–Agent Teams. *Robust Intelligence and Trust in Autonomous Systems*, 167–190. Springer US.
Gillespie, N., and G. Dietz. 2009. "Trust Repair after an Organization-level Failure." *Academy of Management Review* 34 (1): 127–145.
Goodall, N. J. 2014. Machine Ethics and Automated Vehicles. *Lecture Notes in Mobility*, 93–102. Springer International Publishing.
Gottman, J. M. 2005. *The Mathematics of Marriage: Dynamic Nonlinear Models*. Cambridge, MA: MIT Press.
Gottman, J. M., and R. W. Levenson. 1992. "Marital Processes Predictive of Later Dissolution: Behavior, Physiology, and Health." *Journal of Personality and Social Psychology* 63 (2): 221–233.
Graff, G. M. 2017. "How a Dorm Room Minecraft Scam Brought down the Internet." *Wired*, December 13. https://www.wired.com/story/mirai-botnet-minecraft-scam-brought-down-the-internet/.
de Greeff, J., K. Hindriks, M. A. Neerincx, and I. Kruijff-Korbayova 2015. "Human–Robot Teamwork in USAR Environments." *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts – HRI'15 Extended Abstracts*. Portland, Oregon. doi:10.1145/2701973.2702031.
Gurney, J. K. 2013. "Sue My Car Not Me: Products Liability and Accidents Involving Autonomous Vehicles." *Journal of Jaw, Technology & Policy* 2013 (2): 247–277.
Hancock, P. A. 2017. "Imposing Limits on Autonomous Systems." *Ergonomics* 60 (2): 284–291.
Hancock, P. A., D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman. 2011. "A Meta-analysis of Factors Affecting Trust in Human–Robot Interaction." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 53 (5): 517–527.
Haslam, N. 2006. "Dehumanization: An Integrative Review." *Personality and Social Psychology Review* 10 (3): 252–264.
Haslam, N., P. Bain, L. Douge, M. Lee, and B. Bastian. 2005. "More Human than You: Attributing Humanness to Self and Others." *Journal of Personality and Social Psychology* 89 (6): 937–950.
Hayes, C. C., and C. A. Miller. 2016. *Human–Computer Etiquette: Cultural Expectations and the Design Implications They Place on Computers and Technology*. Boca Raton, FL: CRC Press.
Helldin, T., G. Falkman, M. Riveiro, A. Dahlbom, and M. Lebram. 2013. "Transparency of Military Threat Evaluation through Visualizing Uncertainty and System Rationale." *Lecture Notes in Computer Science*, 263–272. Springer Berlin Heidelberg.
Hoff, K. A., and M. Bashir. 2015. "Trust in Automation." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57 (3): 407–434.
Hoffman, R. R., M. Johnson, J. M. Bradshaw, and A. Underbrink. 2013. "Trust in Automation." *IEEE Intelligent Systems* 28 (1): 84–88.
Hoffman, R. R., J. D. Lee, D. D. Woods, N. Shadbolt, J. Miller, and J. M. Bradshaw. 2009. "The Dynamics of Trust in Cyberdomains." *IEEE Intelligent Systems* 24 (6): 5–11.
Humphreys, C., R. Cobb, D. Jacques, and J. Reeger. 2015. "Optimal Mission Path for the Uninhabited Loyal Wingman." *16th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*. Dallas, Texas. doi:10.2514/6.2015-2792.
Hutchins, A. R., M. L. Cummings, M. Draper, and T. Hughes. 2015. "Representing Autonomous Systems' Self-confidence through Competency Boundaries." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 59 (1): 279–283.
Inbar, O., and J. Meyer. 2015. "Manners Matter." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 59 (1): 185–189.
Jamieson, G. A., and G. Skraaning. 2017. "Levels of Automation in Human Factors Models for Automation Design: Why We Might Consider Throwing the Baby out with the Bathwater." *Journal of Cognitive Engineering and Decision Making* 12 (1): 42–49.
Johnson, D., J. Gardner, and J. Wiles. 2004. "Experience as a Moderator of the Media Equation: The Impact of Flattery and Praise." *International Journal of Human–Computer Studies* 61 (3): 237–258.
Johnson, M., J. M. Bradshaw, and P. J. Feltovich. 2017. "Tomorrow's Human–Machine Design Tools: From Levels of Automation to Interdependencies." *Journal of Cognitive Engineering and Decision Making* 12 (1): 77–82.
Jonsson, I.-M., C. Nass, J. Endo, B. Reaves, H. Harris, J. Le Ta, N. Chan, and S. Knapp. 2004. "Don't Blame Me I Am Only the Driver." *Extended Abstracts of the 2004 Conference on Human Factors and Computing Systems – CHI '04*. doi:10.1145/985921.986028.
Jung, M. F. 2017. "Affective Grounding in Human-Robot Interaction." *Proceedings of the 2017 ACM/IEEE International*

Conference on Human–Robot Interaction – HRI '17. Vienna, Austria. doi:10.1145/2909824.3020224.

Jung, M. F., N. Martelaro, and P. J. Hinds. 2015. "Using Robots to Moderate Team Conflict." *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human–Robot Interaction – HRI '15*. Portland, Oregon. doi:10.1145/2696454.2696460.

Kaber, D. B. 2017a. "A Conceptual Framework of Autonomous and Automated Agents." *Theoretical Issues in Ergonomics Science* 28 (3): 1–25.

Kaber, D. B. 2017b. "Issues in Human–Automation Interaction Modeling: Presumptive Aspects of Frameworks of Types and Levels of Automation." *Journal of Cognitive Engineering and Decision Making* 12 (1): 7–24.

Kim, P. H., K. T. Dirks, and C. D. Cooper. 2009. "The Repair of Trust: A Dynamic Bilateral Perspective and Multilevel Conceptualization." *Academy of Management Review* 34 (3): 401–422.

Kim, P. H., K. T. Dirks, C. D. Cooper, and D. L. Ferrin. 2006. "When More Blame is Better than Less: The Implications of Internal Vs. External Attributions for the Repair of Trust after a Competence- Vs. Integrity-based Trust Violation." *Organizational Behavior and Human Decision Processes* 99 (1): 49–65.

Kim, P. H., D. L. Ferrin, C. D. Cooper, and K. T. Dirks. 2004. "Removing the Shadow of Suspicion: The Effects of Apology versus Denial for Repairing Competence- versus Integrity-based Trust Violations." *Journal of Applied Psychology* 89 (1): 104–118.

Kirlik, A. 2017. "Automation and Adaptive Behavior." *Journal of Cognitive Engineering and Decision Making* 12 (1): 70–73.

Klein, G., D. D. Woods, J. M. Bradshaw, R. R. Hoffman, and P. J. Feltovich. 2004. "Ten Challenges for Making Automation a 'Team Player' in Joint Human–Agent Activity." *IEEE Intelligent Systems* 19 (6): 91–95.

Koo, J., J. Kwac, W. Ju, M. Steinert, L. Leifer, and C. Nass. 2014. "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* 9 (4): 269–275.

Kramer, R. M., and R. J. Lewicki. 2010. "Repairing and Enhancing Trust: Approaches to Reducing Organizational Trust Deficits." *The Academy of Management Annals* 4 (1): 245–277.

Kraus, J. M., J. Sturn, J. E. Reiser, and M. Baumann. 2015. "Anthropomorphic Agents, Transparent Automation and Driver Personality." *Adjunct Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications – Automotive UI '15*. doi:10.1145/2809730.2809738.

Kurzweil, R. 2005. *The Singularity is near: When Humans Transcend Biology*. London: Penguin.

Kushner, D. 2013. "The Real Story of Stuxnet." *IEEE Spectrum* 50 (3): 48–53.

Kwiatkowska, M., and M. Lahijanian. 2016. "Social Trust: A Major Challenge for the Future of Autonomous Systems." *AAAI Fall Symposium on Cross-disciplinary Challenges for Autonomous Systems, AAAI Fall Symposium. AAAI*, September. Arlington, VA.

Lee, J. D. 2017. "Perspectives on Automotive Automation and Autonomy." *Journal of Cognitive Engineering and Decision Making* 12 (1): 53–57.

Lee, J. D., and N. Moray. 1992. "Trust, Control Strategies and Allocation of Function in Human–Machine Systems." *Ergonomics* 35 (10): 1243–1270.

Lee, N., J. Kim, E. Kim, and O. Kwon. 2017. "The Influence of Politeness Behavior on User Compliance with Social Robots in a Healthcare Service Setting." *International Journal of Social Robotics* 9 (5): 727–743.

Lewicki, R. J., E. C. Tomlinson, and N. Gillespie. 2006. "Models of Interpersonal Trust Development: Theoretical Approaches, Empirical Evidence, and Future Directions." *Journal of Management* 32 (6): 991–1022.

Lin, P. 2016. "Why Ethics Matters for Autonomous Cars." In *Autonomous Driving*, 69–85. Berlin: Springer.

Long, S. K., N. D. Karpinsky, and J. P. Bliss. 2017. "Trust of Simulated Robotic Peacekeepers among Resident and Expatriate Americans." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 61 (1): 2091–2095.

Lyons, J. B. 2013. "Being Transparent about Transparency: A Model for Human–Robot Interaction." *2013 AAAI Spring Symposium Series*. Arlington, VA. http://www.aaai.org/ocs/index.php/SSS/SSS13/paper/viewPaper/5712.

Madhavan, P., and D. A. Wiegmann. 2004. "A New Look at the Dynamics of Human–Automation Trust: Is Trust in Humans Comparable to Trust in Machines?" *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 48 (3): 581–585.

Madhavan, P., and D. A. Wiegmann. 2007a. "Effects of Information Source, Pedigree, and Reliability on Operator Interaction with Decision Support Systems." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 49 (5): 773–785.

Madhavan, P., and D. A. Wiegmann. 2007b. "Similarities and Differences between Human–Human and Human–Automation Trust: An Integrative Review." *Theoretical Issues in Ergonomics Science* 8 (4): 277–301.

Marinaccio, K., S. Kohn, R. Parasuraman, and E. J. de Visser. 2015. "A Framework for Rebuilding Trust in Social Automation across Health-care Domains." *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care* 4 (1): 201–205.

McCullough, M. E., E. L. Worthington, J. Maxey, and K. C. Rachal. 1997. "Gender in the Context of Supportive and Challenging Religious Counseling Interventions." *Journal of Counseling Psychology* 44 (1): 80–88.

McGuirl, J. M., and N. B. Sarter. 2006. "Supporting Trust Calibration and the Effective Use of Decision Aids by Presenting Dynamic System Confidence Information." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 48 (4): 656–665.

McIntyre, R. M., and E. Salas. 1995. "Measuring and Managing for Team Performance: Emerging Principles from Complex Environments." In *Team Effectiveness and Decision Making in Organizations*, edited by R. Guzzo and E. Salas, 149–203. San Francisco, CA: Jossey-Bass.

McKendrick, R., T. Shaw, E. de Visser, H. Saqer, B. Kidwell, and R. Parasuraman. 2013. "Team Performance in Networked Supervisory Control of Unmanned Air Vehicles." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 56 (3): 463–475.

Mercado, J. E., M. A. Rupp, J. Y. C. Chen, M. J. Barnes, D. Barber, and K. Procci. 2016. "Intelligent Agent Transparency in Human–Agent Teaming for Multi-UxV Management." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 58 (3): 401–415.

Merritt, S. M., and D. R. Ilgen. 2008. "Not All Trust is Created Equal: Dispositional and History-based Trust in Human–Automation

Interactions." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 50 (2): 194–210.
Meyer, J., H. Becker, P. M. Bösch, and K. W. Axhausen. 2017. "Autonomous Vehicles: The Next Jump in Accessibilities?" *Research in Transportation Economics* 62 (Supplement C): 80–91.
Meyer, J., C. Miller, P. A. Hancock, E. J. de Visser, and M. Dorneich. 2016. "Politeness in Machine–Human and Human–Human Interaction." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 60 (1): 279–283.
Meyerson, D., K. E. Weick, and R. M. Kramer. 1996. "Swift Trust and Temporary Groups." In *Trust in Organizations: Frontiers of Theory and Research*, 166–195. Thousand Oaks, CA: Sage.
Miller, C. A. 2017. "The Risks of Discretization: What is Lost in (Even Good) Levels-of-Automation Schemes." *Journal of Cognitive Engineering and Decision Making* 12 (1): 74–76.
Mittu, R., D. Sofge, A. Wagner, and W. F. Lawless. 2016. *Robust Intelligence and Trust in Autonomous Systems*. Boston, MA: Springer.
Morita, P. P., and C. M. Burns. 2012. "Understanding 'Interpersonal Trust' from a Human Factors Perspective: Insights from Situation Awareness and the Lens Model." *Theoretical Issues in Ergonomics Science* 15 (1): 88–110.
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.
Muir, B. M., and N. Moray. 1996. "Trust in Automation. Part II. Experimental Studies of Trust and Human Intervention in a Process Control Simulation." *Ergonomics* 39 (3): 429–460.
Naikar, N. 2017. "Human–Automation Interaction in Self-organizing Sociotechnical Systems." *Journal of Cognitive Engineering and Decision Making* 12 (1): 62–66.
Nakayachi, K., and M. Watabe. 2005. "Restoring Trustworthiness after Adverse Events: The Signaling Effects of Voluntary 'Hostage Posting' on Trust." *Organizational Behavior and Human Decision Processes* 97 (1): 1–17.
Nass, C., B. J. Fogg, and Y. Moon. 1996. "Can Computers be Teammates?" *International Journal of Human–Computer Studies* 45 (6): 669–678.
Nass, C., I.-M. Jonsson, H. Harris, B. Reaves, J. Endo, S. Brave, and L. Takayama. 2005. "Improving Automotive Safety by Pairing Driver Emotion and Car Voice Emotion." *CHI '05 Extended Abstracts on Human Factors in Computing Systems – CHI '05*. Portland, Oregon. doi:10.1145/1056808.1057070.
Nass, C., and K. M. Lee. 2001. "Does Computer-synthesized Speech Manifest Personality? Experimental Tests of Recognition, Similarity-attraction, and Consistency-attraction." *Journal of Experimental Psychology. Applied* 7 (3): 171–181.
Nass, C., Y. Moon, B. J. Fogg, B. Reeves, and C. Dryer. 1995. "Can Computer Personalities be Human Personalities?" *Conference Companion on Human Factors in Computing Systems – CHI '95*. Denver, Colorado. doi:10.1145/223355.223538.
Nass, C., J. Steuer, and E. R. Tauber. 1994. "Computers are Social Actors." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Celebrating Interdependence – CHI '94*. Boston, Massachusetts. doi:10.1145/191666.191703.
National Highway Traffic Safety Administration. 2016. *Federal Automated Vehicles Policy: Accelerating the Next Revolution in Roadway Safety*. http://www.nhtsa.gov/nhtsa/av/pdf/Federal_Automated_Vehicles_Policy.pdf.
Nelson, E. S. 2007. "The Face Symbol: Research Issues and Cartographic Potential." *Cartographica: The International Journal for Geographic Information and Geovisualization* 42 (1): 53–64.
Norman, D. A. 1988. *The Psychology of Everyday Things*. New York: Basic Books.
Onnasch, L., C. D. Wickens, H. Li, and D. Manzey. 2014. "Human Performance Consequences of Stages and Levels of Automation." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 56 (3): 476–488.
Pak, R., N. Fink, M. Price, B. Bass, and L. Sturre. 2012. "Decision Support Aids with Anthropomorphic Characteristics Influence Trust and Performance in Younger and Older Adults." *Ergonomics* 55 (9): 1059–1072.
Pak, R., A. C. McLaughlin, W. Leidheiser, and E. Rovira. 2017. "The Effect of Individual Differences in Working Memory in Older Adults on Performance with Different Degrees of Automated Technology." *Ergonomics* 60 (4): 518–532.
Pak, R., E. Rovira, A. C. McLaughlin, and N. Baldwin. 2016. "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.
Parasuraman, R., and P. Hancock. 1999. "Using Signal Detection Theory and Bayesian Analysis to Design Parameters for Automated Warning Systems." In *Automation, Technology and Human Performance*, edited by Scerbo M. W. Mouloua, 63–67. Mahwah, NJ: Erlbaum.
Parasuraman, R., and C. A. Miller. 2004. "Trust and Etiquette in High-criticality Automated Systems." *Communications of the ACM* 47 (4): 51–55.
Parasuraman, R., and V. Riley. 1997. "Humans and Automation: Use, Misuse, Disuse, Abuse." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 39 (2): 230–253.
Parasuraman, R., T. B. Sheridan, and C. D. Wickens. 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.
Pelegrini Morita, P., P. P. Morita, and C. M. Burns. 2014. "Trust Tokens in Team Development." *Team Performance Management: An International Journal* 20 (1/2): 39–64.
Quinn, D. B., R. Pak, and E. J. de Visser. 2017. "Testing the Efficacy of Human–Human Trust Repair Strategies with Machines." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 61 (1): 1794–1798.
Rice, S., and K. Geels. 2010. "Using System-wide Trust Theory to Make Predictions about Dependence on Four Diagnostic Aids." *The Journal of General Psychology* 137 (4): 362–375.
Riek, L. D., T.-C. Rabinowitch, B. Chakrabarti, and P. Robinson. 2009. "How Anthropomorphism Affects Empathy toward Robots." *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction – HRI '09*. San Diego, CA. doi:10.1145/1514095.1514158.
Riva, P., S. Sacchi, and M. Brambilla. 2015. "Humanizing Machines: Anthropomorphization of Slot Machines Increases Gambling." *Journal of Experimental Psychology. Applied* 21 (4): 313–325.
Robinette, P., A. M. Howard, and A. R. Wagner. 2015. "Timing is Key for Robot Trust Repair." *Lecture Notes in Computer Science*, 574–583. Springer International Publishing.
Robinette, P., A. M. Howard, and A. R. Wagner. 2017. "Effect of Robot Performance on Human–Robot Trust in Time-critical Situations." *IEEE Transactions on Human–Machine Systems*: 1–12.

Rocco, E. 1998. "Trust Breaks down in Electronic Contexts but Can Be Repaired by Some Initial Face-to-Face Contact." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI '98*. Los Angeles, CA. doi:10.1145/274644.274711.

Roemer, N., S. Jones, M. Marino, S. Hyland, and G. Southwood. 2017. "Electric Autonomous Vehicle Case Study Analysis." https://kb.osu.edu/dspace/handle/1811/80713.

Rousseau, D. M., S. B. Sitkin, R. S. Burt, and C. Camerer. 1998. "Not so Different After All: A Cross-discipline View of Trust." *Academy of Management Review* 23 (3): 393–404.

Rovira, E., R. Pak, and A. McLaughlin. 2016. "Effects of Individual Differences in Working Memory on Performance and Trust with Various Degrees of Automation." *Theoretical Issues in Ergonomics Science* 18 (6): 573–591.

Salas, E., and J. A. Cannon-Bowers. 2001. "The Science of Training: A Decade of Progress." *Annual Review of Psychology* 52 (1): 471–499.

Salas, E., N. J. Cooke, and M. A. Rosen. 2008. "On Teams, Teamwork, and Team Performance: Discoveries and Developments." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 50 (3): 540–547.

Salas, E., D. E. Sims, and C. S. Burke. 2005. "Is There a 'Big Five' in Teamwork?" *Small Group Research* 36 (5): 555–599.

Salonen, A. O. 2018. "Passenger's Subjective Traffic Safety, in-Vehicle Security and Emergency Management in the Driverless Shuttle Bus in Finland." *Transport Policy* 61 (Supplement C): 106–110.

Sarter, N. B., and D. D. Woods. 1997. "Team Play with a Powerful and Independent Agent: Operational Experiences and Automation Surprises on the Airbus A-320." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 39 (4): 553–569.

Satterfield, K., C. Baldwin, E. de Visser, and T. Shaw. 2017. "The Influence of Risky Conditions in Trust in Autonomous Systems." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 61 (1): 324–328.

Schaefer, K. E., J. Y. C. Chen, J. L. Szalma, and P. A. Hancock. 2016. "A Meta-analysis of Factors Influencing the Development of Trust in Automation." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 58 (3): 377–400.

Schaefer, K. E., E. R. Straub, J. Y. C. Chen, J. Putney, and A. W. Evans. 2017. "Communicating Intent to Develop Shared Situation Awareness and Engender Trust in Human–Agent Teams." *Cognitive Systems Research* 46: 26–39.

Schilke, O., M. Reimann, and K. S. Cook. 2013. "Effect of Relationship Experience on Trust Recovery following a Breach." *Proceedings of the National Academy of Sciences* 110 (38): 15236–15241.

Schweitzer, M., J. Hershey, and E. Bradlow. 2006. "Promises and Lies: Restoring Violated Trust." *Organizational Behavior and Human Decision Processes* 101 (1): 1–19.

Seeger, A.-M., and A. Heinzl. 2017. "Human versus Machine: Contingency Factors of Anthropomorphism as a Trust-inducing Design Strategy for Conversational Agents." *Lecture Notes in Information Systems and Organisation*, 129–139.

Semigran, H. L., D. M. Levine, S. Nundy, and A. Mehrotra. 2016. "Comparison of Physician and Computer Diagnostic Accuracy." *JAMA Internal Medicine* 176 (12): 1860–1861.

Seo, S. H., K. Griffin, J. E. Young, A. Bunt, S. Prentice, and V. Loureiro-Rodríguez. 2017. "Investigating People's Rapport Building and Hindering Behaviors When Working with a Collaborative Robot." *International Journal of Social Robotics* 10 (1): 147–161.

Sheridan, T. B. 2016. "Human–Robot Interaction." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 58 (4): 525–532.

Sheridan, T. B. 2017. "Comments on 'Issues in Human–Automation Interaction Modeling: Presumptive Aspects of Frameworks of Types and Levels of Automation' by David B. Kaber." *Journal of Cognitive Engineering and Decision Making* 12 (1): 25–28.

Singh, I. L., R. Molloy, and R. Parasuraman. 1993. "Individual Differences in Monitoring Failures of Automation." *The Journal of General Psychology* 120 (3): 357–373.

Slovic, P. 1993. "Perceived Risk, Trust, and Democracy." *Risk Analysis* 13 (6): 675–682.

Slovic, P. 1999. "Trust, Emotion, Sex, Politics, and Science: Surveying the Risk-assessment Battlefield." *Risk Analysis: An Official Publication of the Society for Risk Analysis* 19 (4): 689–701.

Smith, P. J. 2017. "Conceptual Frameworks to Guide Design." *Journal of Cognitive Engineering and Decision Making* 12 (1): 50–52.

Sorkin, R. D., B. H. Kantowitz, and S. C. Kantowitz. 1988. "Likelihood Alarm Displays." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 30 (4): 445–459.

Srinivasan, V., and L. Takayama. 2016. "Help Me Please." *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems – CHI '16*. San Jose, CA. doi:10.1145/2858036.2858217.

Szalma, J. L. 2014. "On the Application of Motivation Theory to Human Factors/Ergonomics." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 56 (8): 1453–1471.

Szalma, J. L., and G. S. Taylor. 2011. "Individual Differences in Response to Automation: The Five Factor Model of Personality." *Journal of Experimental Psychology. Applied* 17 (2): 71–96.

Thielmann, I., and B. E. Hilbig. 2015. "Trust: An Integrative Review from a Person–Situation Perspective." *Review of General Psychology* 19 (3): 249–277.

Tomlinson, E. C., B. R. Dineen, and R. J. Lewicki. 2004. "The Road to Reconciliation: Antecedents of Victim Willingness to Reconcile following a Broken Promise." *Journal of Management* 30 (2): 165–187.

Tomlinson, E. C., and R. C. Mayer. 2009. The Role of Causal Attribution Dimensions in Trust Repair. *Academy of Management Review* 34 (1): 85–104.

Torrey, C., S. R. Fussell, and S. Kiesler. 2013. "How a Robot Should Give Advice." *2013 8th ACM/IEEE International Conference on Human–Robot Interaction (HRI)*. Tokyo, Japan. doi:10.1109/hri.2013.6483599.

Tzeng, J.-Y. 2004. "Toward a More Civilized Design: Studying the Effects of Computers That Apologize." *International Journal of Human-Computer Studies* 61 (3): 319–345.

de Visser, E. J., M. Cohen, A. Freedy, and R. Parasuraman. 2014. "A Design Methodology for Trust Cue Calibration in Cognitive Agents." *Lecture Notes in Computer Science*, 251–262. Springer International Publishing.

de Visser, E. J., S. S. Monfort, K. Goodyear, L. Lu, M. O'Hara, M. R. Lee, R. Parasuraman, and F. Krueger. 2017. "A Little Anthropomorphism Goes a Long Way." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 59 (1): 116–133.

de Visser, E. J., S. S. Monfort, R. McKendrick, M. A. B. Smith, P. E. McKnight, F. Krueger, and R. Parasuraman. 2016. "Almost Human: Anthropomorphism Increases Trust Resilience in Cognitive Agents." *Journal of Experimental Psychology. Applied* 22 (3): 331–349.

de Visser, E. J., R. Pak, and M. A. Neerincx. 2017. "Trust Development and Repair in Human–Robot Teams." *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction – HRI '17*. Vienna, Austria. doi:10.1145/3029798.3038409.

de Visser, E. J., and R. Parasuraman. 2011. "Adaptive Aiding of Human–Robot Teaming." *Journal of Cognitive Engineering and Decision Making* 5 (2): 209–231.

de Visser, E. J., R. Parasuraman, A. Freedy, E. Freedy, & G. Weltman. (2006, October). A Comprehensive Methodology for Assessing Human-Robot Team Performance for Use in Training and Simulation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 50, No. 25, pp. 2639–2643). Los Angeles, CA: SAGE Publications.

Walliser, J. C. 2017. "Social Interactions with Autonomous Agents: Team Perception and Team Development Improve Teamwork Outcomes." PhD, George Mason University.

Walliser, J. C., E. J. de Visser, and T. H. Shaw. 2016. "Application of a System-wide Trust Strategy When Supervising Multiple Autonomous Agents." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 60 (1): 133–137.

Waytz, A., J. Heafner, and N. Epley. 2014. "The Mind in the Machine: Anthropomorphism Increases Trust in an Autonomous Vehicle." *Journal of Experimental Social Psychology* 52: 113–117.

Wickens, C. 2017. "Automation Stages & Levels, 20 Years after." *Journal of Cognitive Engineering and Decision Making* 12 (1): 35–41.

Wiese, E., T. Shaw, D. Lofaro, and C. Baldwin. 2017. "Designing Artificial Agents as Social Companions." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 61 (1): 1604–1608.

Winter, R. 2016. "Using System-wide Trust Theory to Make Predictions about Trust in Transportation Automation." *Journal of Technologies and Human Usability* 12 (2): 1–11.

Woods, D. D. 2016. "The Risks of Autonomy." *Journal of Cognitive Engineering and Decision Making* 10 (2): 131–133.

Woods, D. D., N. Leveson, and E. Hollnagel. 2012. *Resilience Engineering: Concepts and Precepts*. Aldershot, UK: Ashgate Publishing.

Yang, X. J., X. Jessie Yang, V. V. Unhelkar, K. Li, and J. A. Shah. 2017. "Evaluating Effects of User Experience and System Transparency on Trust in Automation." *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction – HRI '17*. doi:10.1145/2909824.3020230.

Zuk, T., and S. Carpendale. 2007. "Visualization of Uncertainty and Reasoning." In *Smart Graphics*, 164–177. Berlin, Heidelberg: Springer.