CHAPTER

16

The role of interdependence in trust

Matthew Johnson and Jeffrey M. Bradshaw

Florida Institute for Human and Machine Cognition, Pensacola, FL, United States

Introduction

As technology continues to take on roles of greater scope and consequence, the issue of trust has moved increasingly to the forefront. Early advances in robotics and automation were typically applied in relatively simple contexts and structured environments where trustworthiness could be straightforwardly assessed and assured (Moorman, Deshpande, & Zaltman, 1993). However, with an increasing number of applications that attempt to cope with complex and uncertain problems once reserved almost exclusively for human judgment, new theories and methods to assess and assure trustworthiness have become imperative. As tragedies traced to defects in the design of complex systems proliferate (e.g. Travis, 2019; Levin & Beene, 2018) it has become clear that traditional approaches to trust fail to address the new risks and reduced transparency of advanced automation and artificial intelligence-based approaches.

Much of the previous work on trust in automation focuses merely on identifying factors that influence people's trust (Hancock et al., 2011; Hoff & Bashir, 2014; Lewandowsky, Mundy, & Tan, 2000; Moorman et al., 1993; Schaefer et al., 2014). Agreement on these factors is useful and sometimes provides helpful information about the importance of situation awareness for automation users. However, knowing factors often does not provide guidance with enough specification detail to influence design decisions that can help avoid such problems in the first place (Hoffman & Deal, 2008).

Though, unfortunately, some researchers still pursue the goal of merely promoting greater trust in automation, we believe that a better research

https://doi.org/10.1016/B978-0-12-819472-0.00016-2

objective is to find ways to help people continuously maintain an *appropriate* level of reliance on technology, taking into account the reliability of the capabilities of the system as it functions at a given time and in a given situation context (Hoffman, Johnson, Bradshaw, & Underbrink, 2013; Lee & See, 2004). The latter objective has led to productive discussions in the research community about how to facilitate reliable "trust calibration." Research on trust calibration seeks ways to enhance an individual's ability to accurately assess the trustworthiness of the technology in different circumstances.

People always arrive at a relationship to some item of technology with some initial bias that leads them to under trust or overtrust its reliability and performance. This initial bias is, of course, outside of the control of those who designed and built the artifact. Fortunately, under normal circumstances, this initial trust assessment will be modified and updated as people use the technology. The trust people gain with the experience of use is what designers and builders can influence. This underscores the point that trust is not a static state, but rather the continuously varying result of a dynamic process (Bradshaw et al., 2004; Hoffman et al., 2009; Lewandowsky et al., 2000; Mayer, Davis, & Schoorman, 1995). For this reason, it is critical to understand the process by which trust evolves (Mayer et al., 1995), so designers and builders can understand how their implementation choices might increase or decrease both the trustworthiness of the technology and also the ability of people to accurately assess that trustworthiness.

In this chapter, we argue that interdependence relationships are the mechanism employed to actively manage the processes that assess trustworthiness of the technology and that enable accurate trust calibration by the people who use it. Trust is relational (Mayer et al., 1995) and interdependence is "the set of complementary relationships that two or more parties rely on to manage required (hard) or opportunistic (soft) dependencies in joint activity" (Johnson et al., 2014). Interdependence relationships constitute the necessary junctures in task-oriented communication and actions among people and machines that make the joint activity in which they are engaged productive (Johnson et al., 2014). Trust is commonly defined as "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor" (Mayer et al., 1995), making the trustor dependent on the trustee. Typically, the relationship of trust among the collaborating parties in nontrivial situations is not a one-way transaction, making the parties not simply dependent but rather mutually *interdependent*. Interactions between humans and machines in mutually interdependent relationships allow them to support one another with respect to current interdependencies. The degree of performance and

reliability in current interactions helps actors accurately calibrate trust with respect to future interactions.

To understand how interdependence relationships are used to actively manage the assessment of trustworthiness and trust calibration, we will extend Mayer et al.'s (1995) model of human–human trust to propose a new model appropriate for groups of humans and machines. Then, we will discuss the role of interdependence in the new model. We will show how the model has allowed us to extend and employ Interdependence Analysis (IA) tables (Johnson, Vignati, & Duran, 2018) as an effective tool for understanding and designing systems capable of actively managing trust through interdependence relationships.

Model of a risk-taking trust relationship

One of the most cited models of trust is that of Mayer et al. (1995). Though developed for examining a single unidirectional human–human trust dyad in an organizational context, its appropriateness for human–machine trust relationships will be argued here. The original model is extended to consider a broader view of activity with refinements to the role of trust. This new model is depicted in Fig. 1.

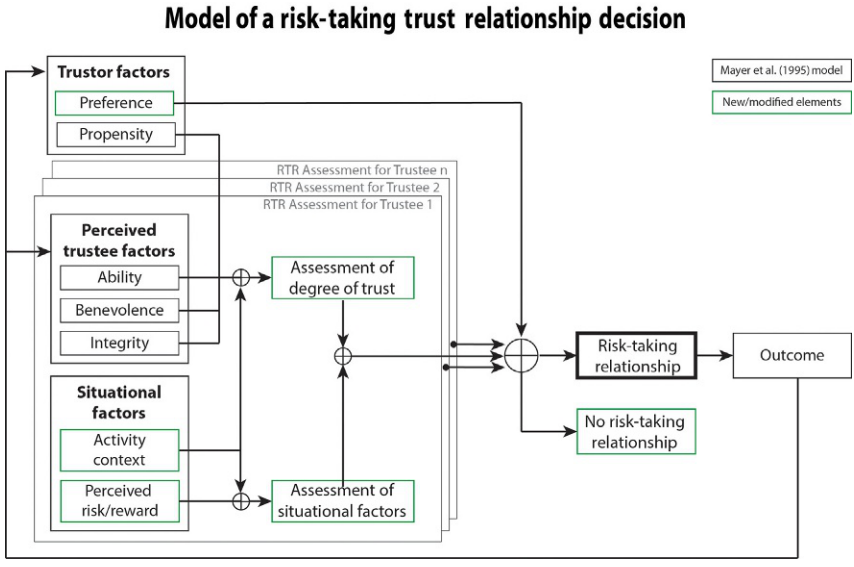


FIG. 1 Model of a risk-taking trust relationship decision, extended and refined from Mayer et al.'s model of trust (1995).

The model proposed by Mayer et al. (1995) is a model of trust, but emphasized the importance of trusting actions taken in the context of a specific risk-taking relationship (RTR). "There is no risk taken in the willingness to be vulnerable (i.e., to trust), but the risk is inherent in the behavioral manifestation of the willingness to be vulnerable. One does not need to risk anything to trust; however, one must take a risk to engage in trusting action" (Mayer et al., 1995, p. 724). Fig. 1 emphasizes a specific RTR decision as a key element in the model, not only because it is the manifestation of trust, but also because it is where the trustor engages in an interdependent relationship with the trustee. The outcome of that decision can strengthen or weaken the assessment of trustworthiness with experience over time. When direct information about the trustee is lacking, people compensate by using what other information they have (Rousseau, Sitkin, Burt, & Camerer, 1998) — for example, relying on the reputation of an institution to which the trustee belongs. However, these initial biases will be informed, modified, and potentially overturned by relational trust based on experience. Experience based on the outcomes of trusting actions, taken in the context of a specific RTR over time, enables accurate trust calibration, generating the objective and subjective bases for justified trust and mistrust. The importance of the RTR is stated by Mayer et al. in this way:

> This relationship-specific boundary condition of our approach is important because a number of authors have dealt with trust for generalized others (e.g., Rotter, 1967) and trust as a social phenomenon (e.g., Lewis & Weigert, 1985). Even though such approaches help provide a general sense of the considerations involved in trust, they do not clarify the relationship between two specific individuals and the reasons why a trustor would trust a trustee. Further, the failure to clearly specify the trustor and the trustee encourages the tendency to change referents and even levels of analysis, which obfuscates the nature of the trust relationship. (Mayer et al., 1995, p. 711).

The model proposed in Fig. 1 remains true to the importance of the relationship. It also extends the idea that context is critical by including additional situational factors and trustor preferences. While the original model was focused on factors affecting trust, the model proposed in Fig. 1 is focused on factors affecting the RTR decision. The RTR will be further examined below, following a description of other elements of the new model.

Characteristics of the trustor

Propensity. The Mayer et al. (1995) model includes the attribute of trustor propensity, which is described as a dispositional willingness to rely on others. People certainly have an innate propensity to either trust or

mistrust technology. It has also been shown that the propensity to trust technology often differs from the propensity to trust people (Madhavan & Wiegmann, 2012).

Preferences. Our model includes an additional factor not found in the original model: a preference. While preference is not a factor in trust assessment, it can play a role in whether an RTR is established or not. One can trust someone and still choose not to engage in an RTR based solely on preference. For example, nobody doubts the efficacy of automatic shifting mechanisms of today's cars, yet some people still choose to manually shift for the pleasure of it. Some people like working with others or using tools and technology, while some get satisfaction from doing it all on their own.

Characteristics of the trustee

The first three trustee factors are identical to the original Mayer et al. (1995) model: ability, benevolence, and integrity. Ability is the skill, competence, expertise of the trustee as perceived by the trustor. Benevolence is the trustor's belief in the trustee's desire to do good on behalf of the trustor. Integrity is the trustor's belief that the trustee adheres to an acceptable set of principles. These factors, determined from a human–human organizational context, seem equally valuable in human–automation contexts.

Commonalities in interpersonal and human–automation trust. It has been asserted that "Human–automation trust and interpersonal trust depend on different attributes. Whereas interpersonal trust can be based on the ability, integrity, or benevolence of a trustee (Mayer et al., 1995), human–automation trust depends on the performance, process, or purpose of an automated system (Lee & Moray, 1992)" (Hoff & Bashir, 2014, p. 413). Contrary to what Hoff and Bashir seem to claim, we see the relevant attributes of the trustee as essentially similar, regardless of whether the trust is interpersonal or directed toward machines. Indeed, two of the authors just cited (Lee and Moray) themselves concluded that "the multidimensional construct of trust developed to describe trust between humans, together with a consideration of the dynamic aspects of trust, can be used to describe trust between humans and machines" (Lee & Moray, 1992, p. 1268).

Ability. Ability plays a similar role in human–human and human–machine trust. Mayer et al. (1995) describe ability as "that group of skills, competencies, and characteristics" of the trustee as perceived by the trustor, this does capture some aspects we would consider important. A more complete definition would be Lee and See's concept of performance defined as "the current and historical operation of the automation and includes characteristics such as reliability, predictability, and ability"

(Lee & See, 2004, p. 59). One key challenge for automation is that humans typically find it easier to estimate the abilities of people than machines. Though people can be trained to make better assessments, they often persist despite themselves in projecting intentionality onto a machine (Nass & Moon, 2000).

Benevolence. While people can certainly hold benevolent feelings toward another, machines do not. There are those who would claim that the designers can build in benevolence, but this is nothing more than wishful mnemonics (McDermott, 1976). Current machines do not have desires, though they may have built-in utility functions. They have no understanding of broader goodness beyond their specific tasks and know little or nothing about the people (potential trustors) with whom they are working. While designers, in general, attempt to make systems "do their best" their attempts are more concerned with ability than benevolence. A notable exception is when designers deliberately try to create software that disables or hijacks machines that don't belong to them. In this sense, software might be seen as "malicious," though it is really the hacker behind the code that has acted with malicious intent. The code is simply "following instructions" that it has been given.

Integrity. In interpersonal relations, integrity is a moral virtue — another wishful mnemonic if it were to be applied literally to machines. However, when Mayer et al. refer to integrity, it means specifically that the trustee adheres to "an acceptable set of principles" (1995, p. 719). This is quite plausible for machines. In some sense, this aligns with Lee and Moray's (1992) concept of purpose, which essentially means carrying out the designer's intent. Depending on how the "acceptable set of principles" is defined, machines could have significantly more integrity than people. To the extent the principles can be formally codified, machines will and indeed must, in ordinary circumstances, meticulously and reliably comply with those principles. While such adherence to principles sounds like a solid win for trusting machines, the problem is in the concrete application of the necessarily abstract principles. This is demonstrated by the well-known example of Isaac Asimov's Laws of Robotics (Asimov, 1950). For instance, what does it mean to "do no harm" when the machine is faced with a zero-sum situation — where choosing to protect one party necessarily causes harm to another and vice versa? While a desirable objective, practical and effective principles for guiding automated behavior are not only difficult to code, but also may be difficult for people to understand.

Even when machines are incapable of benevolence or integrity, people may treat them as if they were acting intentionally (Dennett, 1989). It has been repeatedly demonstrated that people may apply social rules and expectations to computers in a mindless fashion (Nass & Moon, 2000). While designers may be able to exploit people's false projections,

it borders on the unethical to deliberately misrepresent what machines can and can't do except when performing experimental studies.

Situational factors

Risk is a big factor in trust. In the Mayer et al. model, "the perception of risk involves the trustor's belief about likelihoods of gains or losses outside of considerations that involve the relationship with the particular trustee" (1995, p. 726) and this is the same in Fig. 1. What has been added is the consideration of alternatives. While alternatives do not impact whether a trustor trusts a trustee, they do impact the decision to engage in a trusting action (i.e., RTR). Alternatives include both the different ways the trustor could trust and engage with the trustees, as well as the other ways the trustor could accomplish the work (e.g., do it themselves, engage another trustee).

How to choose whether to engage in an RTR

In the original Mayer et al. model, the decision to engage in an RTR was determined by a simple tradeoff between perceptions of trust and risk (Mayer et al., 1995, p. 726). We suggest that additional important considerations may come into play.

The first minor modification is the inclusion of both perceived risk and reward. Trustors will not only evaluate the potential for loss (risk), but also the potential for gain (reward). Greed has often motivated people to take high-risk actions.

The second modification is the addition of Activity Context in situational factors. Trust can only be evaluated with respect to the context in which and the method by which the action is being performed by a specific trustee. This is similar to Lee and See's concept of process, which they described as "Process is the degree to which the automation's algorithms are appropriate for the situation and able to achieve the operator's goals. Process information describes how the automation operates" (Lee & See, 2004, p. 59). For example, a parent may trust their teenager to drive in general but may re-evaluate that trust when considering a more unusual driving challenge like driving across the country or in a busy city. Similarly, parents needing to pick up a package may ask their teenager to choose a longer but more straightforward route than if one of them were doing it themselves. In the model in Fig. 1, the Activity Context attempts to capture these potentially crucial aspects of the decision process.

Activity Context influences both the trustor's assessment of the trustees trustworthiness and the perceived risk/reward. The assessment of trust, driven by the perceived Trustee factors and the Assessment of the

situational factors of activity context and perceived risk/reward are combined into the Trustors assessment of the specific RTR. Trust and risk assessments do not result in a binary decision to engage (trust) or not (mistrust). Hoffman et al. (2009) describe several possible trust relationships that vary in strength. It is not uncommon for people to be uncertain about trust. This is particularly relevant for human–machine trust.

One of the biggest considerations influencing an RTR decision is the options available. People do not consider a single option in isolation from all other options. A trustor might evaluate several alternative RTRs that could achieve the desired outcome, as shown in Fig. 1. Each of these alternatives can have different trustees, different trustworthiness, different activity context, and different risks and rewards. The trustor can consider multiple RTRs as well as the option to do the work themselves when making a choice. All of these options will be considered, as well as a host of additional considerations such as workload, attention demand, and efficiency.

Another often ignored consideration in trust relationships is a personal preference. Simply trusting a trustee is not sufficient to decide to engage in an RTR. The trustor must also prefer to engage in that relationship over other options, including the option to do nothing. Time and limited options might force one to engage with trustees that are not fully trusted. Most people might be willing to engage in a specific RTR even if they are somewhat uncertain about their level of trust, but their preference to do so can vary. Preferences may lead some people to never engage in actions requiring trust.

In short, the trustor will consider the level of trust for a given trustee in a given mode of interaction and situation. A specific RTR may be compared with other acceptable RTRs. Personal preference will help guide the choice of whether or not to engage in a given RTR. The results of an evaluation of one or more RTRs can be used to inform a trustor's decision.

We do not consider our changes to the Mayer et al. model as fundamental. Our model simply broadens its scope so that trust can be more adequately considered in the context of decision making. In the complex realm of human–machine interaction, we find that these additional considerations often play an important role in analysis, experimentation, and deployment.

The role of context

The importance of context has been frequently highlighted by researchers (Bradshaw et al., 2004; Hoffman et al., 2013; Lee & See, 2004; Mayer et al., 1995; Rousseau et al., 1998). Formalizing context has proven a persistent challenge, as has determining relevant context for different problems in different situations.

One basic approach to context is to frame a situation in terms of the *what*, *who*, and *how*. The context of *what* refers to the work that needs to be done or the goal that needs to be achieved. Mayer et al. (1995) state that the question "Do you trust them?" must be qualified with "trust them to do what?" This is Activity Context. Generally speaking, the threshold for trusting a machine to carry your luggage is considerably lower than that required to trust a machine to carry your newborn child. The details of the work and the methods by which it is performed also matter. A parent might trust a teenager to back the car out of the driveway or to drive themselves to school, but may not trust them to drive in the city or at night. To properly analyze and understand trust, modeling the work context in detail (i.e., the *what*) will be essential.

In our basic approach, one must also answer the *who* question. A simple description of the trustee is built into the Mayer et al. model (1995) in terms of ability, benevolence, and integrity. People do establish general feelings of trust toward individuals (e.g. family members, coworkers, teachers). This general feeling must be considered with respect to the details of the work to be done (the *what*). A trustor might trust a family member to drive their car, but they may not trust them to provide accurate medical advice. Conversely, the trustor might trust a doctor to provide medical advice, but not be willing to let the doctor borrow their car. Understanding the contexts of both *what* and *who* are important and the two are related. As noted in the "How to choose whether to engage in an RTR" section, both are needed to have sufficient context for adequately justified decisions. Both are also needed to properly analyze and understand trust.

Our approach to context also requires knowing *how* the work will be done. Part of understanding the context of how is understanding the ability of the trustee (the *who*). What aspects of the work do they consider, and which aspect do they not. It might seem great to some that an automated targeting robot can aim faster and fire more accurately than a human, but this ability loses its potency if the robot's targeting system cannot distinguish friend from foe. This is a critical Activity Context about *how* and is needed when assessing trust in such systems.

Another aspect of *how* is considering the method by which something might be achieved. The inclusion of consideration of alternative methods available to the trustee in the model provides contextual detail necessary for interpreting and assessing whether to engage in a specific RTR. Mayer et al. stated that "the specific consequences of trust will be determined by contextual factors such as the stakes involved, the balance of power in the relationship, the perception of the level of risk, *and the alternatives available to the trustor* [emphasis added]" (1995, pp. 726–727). The various options for methods of accomplishing the work provide a critical context for understanding decisions about whether to engage in an RTR or not. In other words, people consider more than whether they trust the

technology when choosing to rely on it. A trustor might trust the autonomous capabilities of their new Tesla, but trust themselves more in certain traffic situations. The decision might not be based on competence, but on an understanding of context. Or the decision might be purely preferential. The trustor might trust their Tesla to drive the route, but choose to manually drive the scenic route along the coast.

One last contextual consideration with respect to *how* that is important for automation, is the potential interaction available. Does the trustor need to trust it blindly? Can they monitor the automation effectively? Do they have any means to correct or influence the automation after the initial trust engagement? These interactive possibilities are part of the activity context and determine how automation can be engaged in different ways and define a variety of trust processes available.

Trust as a process

One of the key aspects of both the original Mayer et al. model and the proposed one is the feedback loop. When trying to understand trust and risk-taking trust relationships, it is important to remember that trust is a process. While there is a significant consensus on this (Bradshaw et al., 2004; Hoffman et al., 2009; Lewandowsky et al., 2000; Mayer et al., 1995) it is something worth re-emphasizing.

The idea that trust is a process is not merely an observation of human-human dyads or human-automation systems used to support the theory. Trust being a process has serious implications for technology, its design choices, and its acceptance. While an initial impression of trust (ability, benevolence, integrity) might be based on word of mouth, external observation, or blind faith, any long-term use will be based on the outcomes of RTRs. This poses a major challenge to today's technology whose ability to engage effectively in such relationships remains impoverished (Johnson & Vera, 2019). Today's technology relies heavily on reliable performance as the main (sometimes only) feedback. This makes trust in systems brittle and often leaves people ill-informed to make accurate trust assessments in changing contexts. More complex interaction in support of trust would allow for a broader range of trust management not commonly found in today's technology. For example, how does a system convey its abilities or its integrity (i.e., guiding principles)? How can people evaluate a system prior to having to trust it? How do interactions with the system build or erode trust?

Consider the example of a new babysitter. The trustor might ask the babysitter some questions about previous experience to establish credibility. They could also probe ability by asking hypothetical questions such as

"Do you know what to do if there is a fire?" Upon return, conversations about the evening's events can provide feedback that will alter the trustor's perception about the babysitter's ability, benevolence, or integrity.

Machines struggle with all of these options. You rarely can ask them about experience. Only a few systems allow engaging in "what-if" style interaction. The amount of feedback provided is usually quite limited, making it difficult for people to properly calibrate their trust in a given technology. The limited social repertoire of technology has led to a host of issues including automation surprises (Sarter, Woods, & Billings, 1997) and clumsy automation (Wiener & Curry, 1980) to name a few. Many of the limitations are captured in the Ten Challenges of making automation a "Team Player" (Klein, Woods, Bradshaw, Hoffman, & Feltovich, 2004).

Technology's limited capacity to interact not only impedes its effectiveness and trust but can confound research. In many examples comparing interpersonal trust and human-automation trust (Lewandowsky et al., 2000), a human is a "confederate" and no real human-machine interaction is permissible. Thus, for this type of research, the system is forced to be represented as a black box to the participant. Similarly, human–robot interaction research has pointed out that performance is one of the largest influences of trust (Hancock et al., 2011), but this is based on systems with no trust-enabling interaction, leaving performance as the only mechanism for establishing trust.

Human trust over time depends on the development of relational-trust (Rousseau et al., 1998). "Relational trust derives from repeated interactions over time between trustor and trustee. Information available to the trustor from within the relationship itself forms the basis of relational trust" (Rousseau et al., 1998, p. 399). Taking the perspective that trust is a process and that the exchange of information through repeated interactions over time is how appropriate relational trust is developed leads to the role of interdependence.

The role of interdependence

Before explaining the role of interdependence with respect to trust, a summary of its role in human–machine systems, in general, is in order. Coordination is the effective management of dependencies (Malone & Crowston, 1994), or more specifically interdependence relationships (Johnson et al., 2011). Coactive Design, is a unique design approach based on interdependence (Johnson et al., 2011). The goal of Coactive Design is to help designers identify interdependence relationships in a joint activity, so they can design systems that support these relationships, thus enabling designers to achieve the objectives of coordination, collaboration, and

teamwork (Johnson et al., 2014). Instead of considering how to allocate functions, the primary question is how to support interdependence relationships (Johnson, Bradshaw, & Feltovich, 2017).

Coactive Design proposes three essential interdependence relationships: observability, predictability, and directability (Johnson et al., 2014). Observability means making pertinent aspects of one's status, as well as one's knowledge of the team, task, and environment, observable to others. Predictability means that one's actions should be predictable enough that others can reasonably rely on them when considering their own actions. Directability means one's ability to influence the behavior of others and complementarily be influenced by others. These interdependence relationships are essential for effective coordination. They also turn out to be the same foundational relationships that are critical to trust. This was indirectly predicted by Rousseau et al., who noted that "New organizational forms built around the management of interdependence will provide a catalyst for innovative research on trust" (Rousseau et al., 1998, p. 402). So, in what ways do interdependence relationships influence trust?

Trust is developed through interdependence relationships

Interdependence relationships are the mechanism by which relational trust is established. In other words, to establish, develop, and maintain appropriate trust, technology needs to be endowed with appropriate support for interdependence relationships, such as observability, predictability, and directability. The overall goal of any technology should not simply be for people to trust it, but to support the development of appropriate justified trust and mistrust so that people understand when to trust and when not to. The goal should be the establishment of trust that leads to better performance outcomes by the human–machine combination.

Trust, whether between people or between people and machines, is always exploratory. As Hoffman states:

> Active exploration of trusting–relying relationships cannot and should not be aimed at achieving single stable states or maintaining some decontextualized metrical value, but must be aimed at maintaining an appropriate and context-dependent expectation. ***(Hoffman, 2017, p. 157)***

For simple trust relationships with a context suitable for human intuition (e.g. basic physics), nonsocial exploration is certainly feasible. There is no need to have a conversation with a chair to determine if you can stand on it. Some basic incremental loading will provide the feedback you need to decide. It is also unnecessary to understand the inner workings of systems

when how the work gets done is irrelevant. For example, it does not matter how the automatic teller moves money, just that the correct amount of money is presented for retrieval. However, today's technology is pushing into more sophisticated domains with greater complexity and uncertainty. People cannot simply observe an autonomous car drive in a single circumstance and be confident it can handle all driving situations. Doctors will not blindly accept artificial intelligence medical decisions without understanding something about the process behind the decision. Thus, today's sophisticated technology will need to engage in interdependence relationships, like observability, predictability, and directability to foster appropriate trust.

Observability means making pertinent aspects of one's status, as well as one's knowledge of the team, task, and environment, observable to others. This relationship allows a trustor to see or understand pertinent aspects of the trustee. This can and should involve more than just the performance outcome. Consider commercial airline autopilots and their ability to compensate for issues such as asymmetric icing on the wings. These systems silently compensate (Norman, 1990) leaving the pilot unaware of the trouble that is growing. Eventually, they reach their limits and abruptly hand control back to the pilot resulting in a bumpy transfer of controls (Woods & Sarter, 2000). These classic problems with automation are due to a lack of observability. There are many different ways a system can be observable, each with a cost and a benefit. The system could have informed the pilot that it was experiencing icing. If unable to determine that icing was the issue, the system could have informed the pilot that more power than normal was required. The system could have warned the pilot prior to reaching its limits that it was running out of control authority. All of these are examples of making pertinent aspects of the system observable to the human pilot. People have a natural tendency to engage in this type of progress appraisal sharing, even if it is as basic as uttering "something's not right." Machines often lack this basic social competence, even in some of today's most "capable" systems (Johnson & Vera, 2019). A review of Lee and See's design considerations for how to make automation trustable shows that a large number are examples of observability including:

- Show the past performance of the automation.
- Show the process and algorithms of the automation by revealing intermediate results in a way that is comprehensible to the operators.
- Simplify the algorithms and operation of the automation to make it more understandable.
- Show the purpose of the automation, design basis, and range of applications in a way that relates to the users' goals. (Lee & See, 2004, p. 74)

Predictability means that one's actions should be predictable enough that others can reasonably rely on them when considering their own actions. This relationship allows the trustor to establish expectations used to evaluate performance outcomes. Predictability has been a long-standing cornerstone of trust (Lee & See, 2004; Muir & Moray, 1996; Rempel & Holmes, 1985). It is unnecessary to explain why being predictable helps trust and being unpredictable hurts trust. However, it is useful to note the difference between predictable behavior (performance outcome) and supporting a predictable relationship (interaction). Again, progress appraisals play a key role (Feltovich, Bradshaw, Clancey, Johnson, & Bunch, 2008). Statements like "I am running late" or "I am getting tired" help the trustor adjust their expectations and avoid future trust violations from performance outcome deviations. Other means to convey predictability include complying with established norms and being deterministic not only in the outcome but also in how work is accomplished. Even failure can be more acceptable if it is predictable (Muir & Moray, 1996).

Directability means one's ability to influence the behavior of others and complementarily be influenced by others. This relationship allows a trustor to bound or obligates the behavior of a trustee (Bradshaw et al., 2004). While observability (sometimes referred to as transparency) (Chen et al., 2014; Lyons, Wortham, Theodorou, & Bryson, 2013; Yang, Unhelkar, Li, & Shah, 2017) and predictability (Hoff & Bashir, 2014; Lee & See, 2004; Schaefer et al., 2014) both have significant bodies of associated research connecting them to trust, directability has received much less attention. Often it is the ability to direct or influence a trustee that allows a trustor to take the initial steps of partial trust.

> Through experience in seeing the results of providing outside direction to the subject in order to avoid or to recover from failure (whether such failure is inadvertent or intentional) the observer also has an opportunity to learn something about the subject's disposition for compliance: proving the technology to see whether it will do all things that it is commanded. ***(Bradshaw et al., 2004, pp. 19–20)***

To be directable, the trustee needs to support receiving direction from the trustor and must also possess a "disposition for compliance." From an automation perspective, the machine must have a human-usable interface for providing direction and the algorithms must support ingestion and incorporation of the directions. For example, early driving direction software would provide you the "best" route and the user had no say in the results. Now such software offers three routes allowing the user to choose or even drag the route to adjust portions of the result. A trustor's ability to quickly and reliably bring a trustee into compliance can play a role in their risk assessment of the given alternative. Consider

what a driving instructor might be willing to trust a student driver to do if they are using a car in which the instructor has a complete set of duplicate controls, versus a car with only a duplicate brake, versus a car with no controls for the instructor. These alternatives will certainly affect the instructor's willingness to engage in an RTR.

Current deep learning approaches are an example of how a lack of support for interdependence can inhibit trust. Many of today's machine learning approaches fail to support observability, predictability, and directability. People cannot see what the state of the system is or what decisions are based upon. Making these systems more observable is a current area of research (Zeiler & Fergus, 2014). These approaches are notoriously unpredictable. They often provide odd and unpredictable results (Goodfellow, Shlens, & Szegedy, 2014) and demonstrations of their unpredictable brittleness, such as Tesla vehicles being tricked by a few stickers,[a] have become commonplace. These systems provide no indication of how or when they will fail. The design of these systems inhibits directability. They typically do not have interfaces for user input, only data input. Additionally, their algorithms are not designed to support direct input making compliance difficult or even impossible. While today's machine learning approaches are demonstrating amazing leaps in competence (ability), they struggle to support interdependence making trust and acceptance challenging. This is evidenced by recent research programs like the Defense Advanced Research Projects Agency (DARPA) Explainable Artificial Intelligence (XAI) program, which notes that "the effectiveness of these systems is limited by the machine's current inability to explain their decisions and actions to human users.[b]".

Support of interdependence can be achieved through a variety of means. Observability can be achieved by explicit direct communication, like stating "I am tired," or through behavioral displays (Feltovich, Bradshaw, Jeffers, Suri, & Uszok, 2004) such as sighing or body posture. Similarly, predictability can be stating one's intentions verbally or making actions legible (Dragan, Lee, & Srinivasa, 2013) so that one's intentions are clear. Directability can be achieved through commands, suggestions ("Have you tried jiggling it"), progress appraisals ("I'm running late"), warnings ("Watch your step"), helpful adjuncts ("Do you want me to carry that for you?"), and observations about relevant unexpected events ("It has started to rain"). All of these can play a role in allowing the trustee to convey trust signatures to the trustor. This is not a unidirectional process, because the trustor "can not only be made aware of trust and mistrust

[a]https://www.newscientist.com/article/2198325-teslas-autopilot-tricked-into-driving-on-the-wrong-side-of-the-road/

[b]www.darpa.mil/program/explainable-artificial-intelligence.

signatures but can also actively probe the technology (probing the world through the technology) to test hypotheses about trust, and then use the results to adjust subsequent human-machine activities (that is, reliance)" (Hoffman et al., 2013, p. 87).

In summary, the role of interdependence relationships is to support the active and continuous exploration of trust between a trustor and a trustee to ensure trustor assessments are appropriate for achieving the best outcomes possible.

Interdependence relationships enable partial trust

Trust is not all or nothing. There are many ways to trust and many types of trust relationships (Hoffman et al., 2009). Support for interdependence plays a role in which intermediary forms of trust are acceptable and under what conditions. "Degrees of interdependence actually alter the form trust may take" (Rousseau et al., 1998, p. 395). Initially, a parent may trust a child to cross a busy street only while holding their hand. Over time, the parent may be content to watch from the side of the road. Eventually, they will have complete trust. Along the way, support for interdependence (e.g. walking while holding hands, observing the child checking for traffic, observing proper decision making, predictable performance) will enable the establishment and development of trust.

Interdependence relationships help resolve the uncertain

Because interdependence enables partial trust, it helps resolve trust uncertainty. "The need for trusting behavior often arises while there is still a lack of data regarding some of the three factors" (Mayer et al., 1995, p. 730). People will often engage in RTRs to actively explore the suitability of RTRs. They use the feedback from initial engagements to inform future ones, thus reducing uncertainty through experience.

Interdependence relationships encourage opportunistic trust

Trust is not always about required dependence (i.e. a hard constraint). Rousseau et al. stated that "The second necessary condition of trust is interdependence, where the interests of one party *cannot* [emphasis added] be achieved without reliance upon another" (1998, p. 395). However, this is not the only option. There are many examples of trust where the trustor is perfectly capable of achieving the goal without reliance (e.g. commercial airline autopilots). These soft constraint examples of trust highlight the importance of consideration for both

other alternatives and trustor preference when trying to understand RTRs.

Interdependence relationships help manage the complexities of trust

Trust is complex and multidimensional. "The human's stance toward the machine is always some mixture of justified and unjustified trusting and justified and unjustified mistrusting…multiple trusting relations exist simultaneously" (Hoffman, 2017, pp. 152–153). Because of this, there is a need to leverage interdependence relationships to both establish justified trust (i.e. avoid under-reliance) and justified mistrust (avoid over-reliance). Additionally, interdependence relationships can be used to recognize and respond through interaction to counter both positive and negative trust misalignments (avoid unjustified trust and unjustified mistrust). Trusting is an evolving phenomenon. The basis of trust changes as the relationship progresses (Rempel & Holmes, 1985). "A more complete understanding of trust would come from consideration of its evolution within a relationship" (Mayer et al., 1995, p. 727).

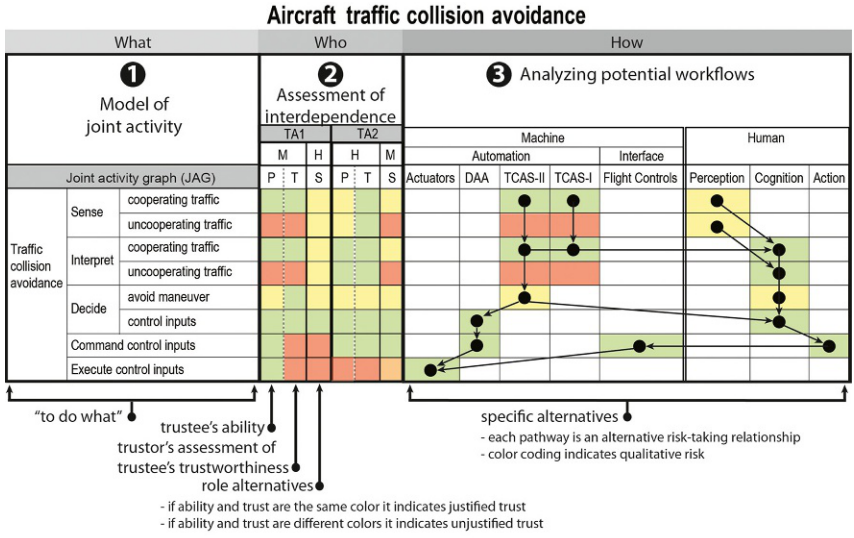
Applying interdependence analysis to trust

Having a model is informative and can provide guidance, but may fall short of influencing design (Hoffman & Deal, 2008). Developers, who are by and large not experts in trust, may have a difficult time interpreting the model and applying it to their design problem. What they need is formative design tools (Johnson, Bradshaw, and Feltovich, 2017) that also help them account for trust.

Coactive Design proposed a design tool called the IA table (Johnson, Bradshaw, et al., 2017; Johnson et al., 2014, 2018). The purpose of IA is understanding how people and automation can effectively team by identifying and providing insight into the potential interdependence relationships used to support one another throughout an activity. This tool already supports many of the key aspects of the proposed trust model and can easily be extended to consider trust and emphasize how interdependence relationships support trust management.

The interdependence analysis table

The IA tool is in the form of a table, as shown in Fig. 2. For discussion purposes, the table presented is populated with a simplified analysis of the aircraft traffic collision avoidance problem. In short, how do you



Color key for team member role alternative capability assessment		
Performer's ability (P)	**Trustor's trust in performer (T)**	**Supporting team members (S)**
I can do it all	Trust	My assistance could improve efficiency
I can do it all but my reliability is < 100%	Circumscribed trust	My assistance could improve reliability
I can contribute but need assistance	Contingent trust	My assistance is required
I cannot do it	Mistrust	I cannot provide assistance
Not applicable /not significant	Uncertain	Not applicable /not significant

TCAS (Traffic Collision Avoidance System) detects and tracks cooperative (transponder equipped) aircraft
TCAS-I provides bearing and altitude and generates collision warnings, but does not offer any suggested remedy
TCAS-II has all of TCAS I, but also offers the pilot direct, vocalized instructions to avoid danger (vertically only)
DAA (Detect and Avoid e.g. ACAS) detection, tracking and avoidance of aircraft

FIG. 2 How the interdependence analysis table relates to the proposed model of trust.

ensure that the aircraft does not collide with other air traffic? The table has three main sections: (1) model of joint activity (the what), (2) assessment of potential interdependence (the who), and (3) analysis of potential workflows (the how).

The what

The first section models the work. While there are many interesting nuances about how to model joint work (Johnson et al., 2018), for this discussion the model of work is the answer to the question "trust it to do *what*?" This provides the activity context. At this point in the analysis, the designer is not concerned with either the *who* (trustee) or the *how*, both of which are needed for a full trust assessment of alternatives. For the aircraft traffic collision avoidance, the work involves sensing traffic, interpreting if it will collide, deciding how to avoid it, and executing the avoidance maneuver.

The who

The second section captures the *who* and provides a systematic way to assesses potential interdependence. It involves enumerating the viable team role alternatives, assessing the ability to perform the work, assessing the ability to support another team member as they perform the work, identifying potential interdependencies, and then determining the requirements to support the interdependence relationships of interest. The team role alternatives capture the set of entities that can potentially participate in the work. The example in Fig. 2 depicts two alternatives. Alternative one (TA1) is the machine (M), in other words, automation, as the performer with the potential for human pilot (H) assistance. The second alternative (TA2) is the human as the performer with the potential for automated assistance.

The first column in an alternative is defined as the performer (P), meaning the entity doing the work. To accommodate a trust assessment, the IA table has been extended with an additional column under performer labeled trustworthiness (T). This is the trustor's subjective assessment of the given performer's trustworthiness in performing the specific aspect of work it aligns within section 1. The remaining columns in an alternative are supporting members (S) assisting the performer. The assessments are generally qualitative assessments captured using a color-coding scheme explained in Fig. 2, though it can be supplemented with empirical quantitative data when available. The color-coding is different for each type of column.

The color-coding of the performer (P) is an assessment of the capacity to perform. This is typically the designer's assessment of the ability of the performer who is the trustee in the relationship. Green means reliable, yellow is less than perfect reliability, orange requires assistance, and red means no ability.

The color-coding for the newly added trustworthiness column (T) reflects the trustor's subjective trustworthiness assessment of the given performer (trustee) to perform the specified aspect of the work. By connecting trust to work we can improve the specificity, which refers to the degree to which trust is associated with a particular component or aspect of the trustee (Lee & See, 2004). Green means the trustee is trusted. Yellow indicates circumscribed trust (Hoffman et al., 2009). This is a more limited trust that might vary with time or activity context. Orange indicates contingent trust (Hoffman et al., 2009). This is an even more limited trust that depends on circumstances. Such trust might demand additional monitoring or progress appraisal checkpoints. Red is mistrust and indicates that the trustor is unlikely to engage in an RTR. Gray is uncertain. The trustor might choose to explore trust by engaging in an RTR, or they may be hesitant to do so depending on their propensity.

The color-coding for the supporting team member column (S) is an assessment of that team member's potential to support the performer for the activity specified by the row. The color red indicates no potential for interdependence, thus the independent operation is the only viable option for the task. Orange indicates a hard constraint, such as providing supplemental lifting capacity when objects are too heavy. Another example of orange is when a machine needs human authorization to perform the activity. Yellow is used to represent improvements to reliability. For example, a human could provide recognition assistance to a robot and increase the reliability in identifying coffee mugs. Green is used to indicate assistance that may improve efficiency. For example, a robot may be able to determine the shortest route much faster than a human or could assist in cleaning up a room. The supporting team member columns are used to identify interdependence requirements (i.e., observability, predictability, directability) needed to support joint activity, but these are also the type of relationships used to calibrate trust.

The purpose of the color-coding is to help identify important design issues with respect to human–machine teaming. The colors and relationships between colors help characterize the design (Johnson et al., 2014). The performer column colors help identify potential brittleness and hard constraints. The supporting team member columns help identify both hard and soft interdependencies (Johnson et al., 2011). With respect to trust, the colors can not only help assess the type of trust but also the correspondence between a person's trust in the automation and the automation's capabilities sometimes referred to as trust calibration (Lee & Moray, 1994). For example, in the scenario in Fig. 2, the Traffic Collision Avoidance System (TCAS[c]) is used to sense traffic. TCAS can only detect cooperating traffic with special equipment. Aircraft without such equipment cannot be sensed and is referred to as uncooperating traffic. Accordingly, in Fig. 2, the machine has the ability to sense cooperating traffic (TA1 column P is green) and the trustor trusts the system (TA1 column T is green). This is justified trust. The machine does not have the ability to sense uncooperating traffic (TA1 column P is red) and the trustor does not trust the system (TA1 column T is red). This is justified mistrust. According to the example, the system is less than 100% reliable at deciding on the proper avoidance maneuver (TA1 column P is yellow), but the trustor trusts the system (TA1 column T is green). This is unjustified trust. Once the decision is made, the system is perfectly capable of executing the avoidance maneuver (TA1 column P is green) yet the trustor does

[c]You would not normally have TCASI and TCAS II in the same system. This example includes both just to demonstrate how you can capture different capabilities. Additionally, TCAS and DAA are often separate systems, but they are shown working together for similar explanatory purposes.

not trust the system to do so (TA1 column T is red). This is unjustified mistrust. These are just a few of the possibilities of how to capture trust in the context of a specific trustee and a specific activity.

The how

The third section connects the theoretical understanding of work to physical instantiation in a specific system. Across the top of the workflow section in Fig. 2, there are column headings for each algorithm, interface element, or human capability used to accomplish the task. Below each heading is a black dot to indicate where in the activity that particular component has a role. The dots are connected with arrows to indicate potential workflows to accomplish the goal. The resulting graph structure is a visual description of all existing and potential workflows, in other words, the alternatives.

The color-coding in the workflow section shown in Fig. 2 is a copy of the performer's ability color-coding. This represents the risk of taking a specific pathway (alternative). For example, the risk of a fully automated response to traffic includes the risk of not seeing uncooperating traffic (red indicates certain to fail) and the risk of making the wrong avoidance maneuver decision (yellow indicates only a possibility of failing). Now consider the fully manual pathway. There is a risk that the human might not see the traffic due to lack of attention (yellow for sensing). However, the human is capable of sensing uncooperating traffic. People, like machines, have the potential to make an incorrect decision about how to avoid traffic (yellow). The IA table allows for risk assessment for specific alternatives and enables the designer to see all of the risks in context, not simply a general risk for automating traffic avoidance as a whole. It also allows comparison of risk across alternatives.

So which option is better: fully automated or manual? The answer, as usual, is neither. The best answer is a combination of the two (Johnson et al., 2017; Johnson & Vera, 2019). The workflows in Fig. 2 show some limited support for interdependence as horizontal arrows that cross between the human and machine sections. In the interpretation activity, TCAS can alert the human about traffic. TCAS-II can provide the pilot with a recommended avoidance maneuver (e.g. "pull up"). These interdependence relationships, alerting (observability), and recommending (directability) enable additional alternatives. The human pilot can be supplemented by the automated solution to potentially outperform either alone. Supplementation means the human can partially mitigate the automation's lack of ability to sense uncooperating traffic. It also means the automation can partially mitigate the human's limited attention.

Although not shown in Fig. 2, the workflow colors could also be based on trust (T column of the performer). While using the ability color-coding

speaks to potential performance outcomes, using the trust color-coding speaks to the potential reliance. It is irrelevant if a system has the potential for high-performance outcomes if the trustor does not trust the system and never relies on it. In the example shown in Fig. 2, the trustor (human pilot) does not trust the system to command the control inputs. This means that no matter how good the automated solution is, the pilot would not rely on it. However, the pilot does trust the system in other ways making both the manual and the human–automation teaming options viable.

The IA table, with small modifications presented here, supports the analysis of human–machine trust based on the proposed model shown in Fig. 1. It provides activity context, directly addresses the trustee's ability, speaks to RTR alternatives, risk, and trust, and allows for distinctions about the different types of trust. The IA table provides a detailed level of context necessary for proper consideration of trust. Additionally, the IA table can help understand and design support for interdependencies (its original purpose) with respect to not only joint activity but also the process of active trust management. In sum, it helps consider all of the major factors influencing engagement in risk-taking trust relationships.

The value of IA is that it provides a detailed contextual lens for interpreting trust. It enables designers, developers, or analysts to capture the *what*, *who*, and *how* contextual aspects of trust in a systematic way. The ability to interpret trust at a finer resolution provides the potential for improved measurement of trust which in turn has the potential to improve predictions of trust. IA also provides a means to compare the risks of different alternatives available which can potentially improve predictions about choices to engage in different RTRs. The contextual detail also enables engineers to identify weaknesses in their design and develop engineering advances aimed at improving active trust management.

Conclusion

This work has proposed a new model of risk-taking trust relationships that is an extension and refinement of Mayer et al.'s (1995) model of trust. It combines the original trustee factors and trustor propensity with considerations for activity context, perceived risk/reward, other RTRs, and trustor preference. The role of interdependence relationships is to support active and continuous exploration of trust between a trustor and a trustee to ensure trustor assessments are appropriate for achieving the best outcomes possible. IA provides a contextual lens for interpreting trust. Future sophisticated technology will need to engage in interdependence relationships, like observability, predictability, and directability to foster appropriate trust calibration. The IA table is ideally suited to the analysis of

human–automation trust and aligns well with the proposed model. It can be an effective tool for understanding and designing systems capable of actively managing trust through interdependence relationships.

References

Asimov, I. (1950). *I Robot*. New York: Bantam Dell.

Bradshaw, J. M., Jung, H., Kulkarni, S., Johnson, M., Feltovich, P., Allen, J. , et al. (2004). Toward trustworthy adjustable autonomy in KAoS. In: *Trusting agents for trusting electronic societies*. https://doi.org/10.1007/11532095_2.

Chen, J. Y. C., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. J. (2014). *Situation awareness-based agent transparency*: (pp. 1–29). US Army Research Laboratory. April, Retrieved from (2014). https://apps.dtic.mil/docs/citations/ADA600351.

Dennett, D. C. (1989). *The intentional stance*. Retrieved from *(1989)*. https://books.google.com/books?hl=en&lr=&id=Qbvkja-J9iQC&oi=fnd&pg=PP11&dq=Dennett,+Daniel+C.+The+Intentional+Stance+&ots=73gnKNk1M0&sig=Q-QcDMDZ_6L340s9Uun9Rv2B85M.

Dragan, A. D., Lee, K. C. T., & Srinivasa, S. S. (2013). Legibility and predictability of robot motion. In: *ACM/IEEE international conference on human-robot interaction, 1*, pp. 301–308. https://doi.org/10.1109/HRI.2013.6483603.

Feltovich, P. J., Bradshaw, J. M., Clancey, W. J., Johnson, M., & Bunch, L. (2008). Progress appraisal as a challenging element of coordination in human and machine joint activity. *Lecture Notes in Computer Science*, *4995 LNAI*, 124–141. (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) (2008). https://doi.org/10.1007/978-3-540-87654-0_6.

Feltovich, P. J., Bradshaw, J. M., Jeffers, R., Suri, N., & Uszok, A. (2004). Social order and adaptability in animal and human cultures as analogues for agent communities: Toward a policy-based approach. A. Omacini, P. Petta, & J. Pitt (Eds.), *Engineering societies in the agents world IV* (pp. 21–48). Vol. Lecture(pp. 21–48). Heidelberg, Germany: Springer.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). *Explaining and harnessing adversarial examples*. https://doi.org/10.1109/CVPR.2015.7298594.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517–527. https://doi.org/10.1177/0018720811417254.

Hoff, K. A., & Bashir, M. (2014). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570.

Hoffman, R. R. (2017). A taxonomy of emergent trusting in the human-machine relationship. In *Cognitive systems engineering: The future for a changing world* (pp. 137–164). Boca Raton, FL: CRC Press, Taylor & Francis. https://doi.org/10.1201/9781315572529.

Hoffman, R. R., & Deal, S. V. (2008). Influencing versus informing design, part 1: A gap analysis. *IEEE Intelligent Systems*, *23*(5), 78–81.

Hoffman, R. R., Johnson, M., Bradshaw, J. M., & Underbrink, A. (2013). Trust in automation. *IEEE Intelligent Systems*. *28*(1). https://doi.org/10.1109/MIS.2013.24.

Hoffman, R. R., Lee, J. D., Woods, D. D., Shadbolt, N., Miller, J., & Bradshaw, J. M. (2009). The dynamics of trust in cyberdomains. *IEEE Intelligent Systems*, *24*(6), 5–11. Retrieved from (2009). http://eprints.ecs.soton.ac.uk/20372/.

Johnson, M., Bradshaw, J. M., & Feltovich, P. J. (2017). Tomorrow's human–machine design tools: From levels of automation to interdependencies. *Journal of Cognitive Engineering and Decision Making*. https://doi.org/10.1177/1555343417736462. 155534341773646.

Johnson, M., Bradshaw, J., Feltovich, P., Jonker, C., van Riemsdijk, B., & Sierhuis, M. (2011). The fundamental principle of coactive design: Interdependence must shape autonomy. M. De Vos, N. Fornara, J. Pitt, & G. Vouros (Eds.), *Coordination, organizations, institutions, and norms in agent systems VI* (Vol. 6541, pp. 172–191). https://doi.org/10.1007/978-3-642-21268-0_10.

Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C. M., van Riemsdijk, B. M., & Sierhuis, M. (2014). Coactive design: Designing support for interdependence in joint activity. *Journal of Human-Robot Interaction*, *3*(1), 43–69.

Johnson, M., Shrewsbury, B., Bertrand, S., Calvert, D., Wu, T., Duran, D. , et al. (2017). Team IHMC's lessons learned from the DARPA robotics challenge: Finding data in the rubble. *Journal of Field Robotics*, *34*(2), 241–261. https://doi.org/10.1002/rob.21674.

Johnson, M., & Vera, A. H. (2019). No AI is an island: The case for teaming intelligence. *AI Magazine, Spring*, 16–28.

Johnson, M., Vignati, M., & Duran, D. (2018). Understanding human-autonomy teaming through interdependence analysis. In: *Symposium on human autonomy teaming*.

Klein, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., & Feltovich, P. J. (2004). Ten challenges for making automation a "team player" in joint human-agent activity. *IEEE Intelligent Systems*, *19*(6), 91–95. https://doi.org/10.1109/MIS.2004.74.

Lee, J. D., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243–1270. Retrieved from (1992). https://www.tandfonline.com/doi/abs/10.1080/00140139208967392.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, *40*(1), 153–184. Retrieved from (1994). https://pdfs.semanticscholar.org/42e4/39534a7952610884b183d123e8a689ac8684.pdf.

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

Levin, A., & Beene, R. (2018). *Tesla's autopilot found partly to blame for 2018 crash on the 405. Retrieved from 4 January 2020, from Los Angeles times website(2018).* https://www.latimes.com/business/story/2019-09-04/tesla-autopilot-is-found-partly-to-blame-for-2018-freeway-crash.

Lewandowsky, S., Mundy, M., & Tan, G. P. (2000). The dynamics of trust: Comparing humans to automation. *Journal of Experimental Psychology Applied*, *6*(2), 104–123. Retrieved from (2000). https://psycnet.apa.org/journals/xap/6/2/104/.

Lewis, J. D., & Weigert, A. (1985). Trust as a social reality source: Social forces. *Social Forces*, *63*(4), 967–985. https://doi.org/10.1093/sf/63.4.967.

Lyons, J. B., Wortham, R. H., Theodorou, A., & Bryson, J. J. (2013). Being transparent about transparency: A model for human-robot interaction. In: *Proceedings of AAAI spring symposium on trust in autonomous systems*, pp. 48–53. https://doi.org/10.1021/ja00880a025.

Madhavan, P., & Wiegmann, D. A. (2012). 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. https://doi.org/10.1177/154193120404800365.

Malone, T. W., & Crowston, K. (1994). The interdisciplinary study of coordination. *ACM Computing Surveys*, *26*(1), 87–119. https://doi.org/10.1145/174666.174668.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.5465/amr.1995.9508080335.

McDermott, D. (1976). Artificial intelligence meets natural stupidity. *ACM SIGART Bulletin*, Retrieved from (1976). http://dl.acm.org/citation.cfm?id=1045340.

Moorman, C., Deshpande, R., & Zaltman, G. (1993). Factors affecting trust in market research relationships. *Journal of Marketing*, *57*(1), 81–101. Retrieved from (1993). https://journals.sagepub.com/doi/abs/10.1177/002224299305700106.

Muir, B. M., & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, *39*(3), 429–460. Retrieved from (1996). https://www.researchgate.net/profile/Neville_Moray/publication/14354918_Trust_in_automation_Part_II_Experimental_studies_of_trust_and_human_intervention_in_a_process_control_simulation/links/58219d0d08ae40da2cb777b9.pdf.

Nass, C., & Moon, Y. (2000). Mindfulness theory and social issues – machines and mindlessness – social responses to computers. *Journal of Social Issues*, *2000*(1), 81–104. Retrieved from (2000). http://www.coli.uni-saarland.de/courses/agentinteraction/contents/papers/Nass00.pdf.

Norman, D. (1990). The "problem" of automation. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences*, *327*(1241), 585–593. Retrieved from(1990). https://pdfs.semanticscholar.org/e7e9/a8ddc88c30bcd408805dab80cba0052f97b9.pdf.

Rempel, J., & Holmes, J. (1985). Trust in close relationships. *Journal of Personality and Social Psychology*, *49*(1), 95–112. Retrieved from(1985). https://www.researchgate.net/profile/John_Holmes9/publication/232554295_Trust_in_close_relationships_J_Pers_Soc_Psychol/links/0046352cdcd5694aee000000/Trust-in-close-relationships-J-Pers-Soc-Psychol.pdf.

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality*, *35*(4), 651–665. https://doi.org/10.1111/j.1467-6494.1967.tb01454.x.

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

Sarter, N. B., Woods, D. D., & Billings, C. E. (1997). Automation surprises. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (pp. 1926–1943). (2nd ed.). Wiley.

Schaefer, K. E., Billings, D. R., Szalma, J. L., Adams, J. K., Sanders, T. L., Chen, J. Y. C. , et al. (2014). A meta-analysis of factors influencing the development of Trust in Automation: Implications for human-robot interaction. In *(No. ARL-TR-6984). Army Research Lab Aberdeen proving ground md human research and engineering directorate*. https://doi.org/10.1177/0018720816634228.

Travis, G. (2019). How the Boeing 737 max disaster looks to a software developer-IEEE Spectrum. *IEEE Spectrum*, Retrieved from(2019). https://spectrum.ieee.org/aerospace/aviation/how-the-boeing-737-max-disaster-looks-to-a-software-developer.

Wiener, E. L., & Curry, R. E. (1980). Flight-deck automation: Promises and problems. *Ergonomics*, *23*(10), 995–1011. https://doi.org/10.1080/00140138008924809.

Woods, D. D., & Sarter, N. B. (2000). *Learning from automation surprises and "going sour" accidents. In Cognitive engineering in the aviation domain* (pp. 327–353). Retrieved from(2000). https://www.researchgate.net/profile/David_Woods11/publication/268275060_Learning_from_Automation_Surprises_and_Going_Sour_Accidents/links/54dcc7620cf282895a3b22a3.pdf.

Yang, X. J., Unhelkar, V. V., Li, K., & Shah, J. A. (2017). Evaluating effects of user experience and system transparency on trust in automation. In: *HRI '17: Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*, pp. 408–416. https://doi.org/10.1145/2909824.3020230.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In: *European conference on computer vision*, Cham: Springer, pp. 818–833.