Special issue: Trust and social capital in teams and organizations

Numéro spécial: Confiance et capital social dans les équipes et les organisations

SSI

Markus Schöbel

Trust in high-reliability organizations

Abstract. *The present article aims to highlight the effects of trust on safety performance in high-reliability organizations (HROs) like nuclear power plants, chemical plants or hospital emergency departments. The author claims that not only beneficial but also detrimental effects have to be considered in the analysis of trust within these socio-technical systems. Potential safety outcomes of trusting behavior are discussed in the light of two types of interaction underlying task management in HROs: trust in human interactions vs. trust in human–system interaction. Trust is further specified according to the constraints and requirements that may interfere with the beneficial role of trusting behavior. In particular, three distinct types of trust beliefs moderating the effect of trust on safety performance are addressed: beliefs based on shared values and norms, institution-based beliefs, and beliefs based on system reliability. Finally, the author highlights organizational factors that emerge as crucial for the development and maintenance of safe work settings in which the beneficial aspects of trust are brought to bear.*

Key words. *High-reliability organizations – Human interaction – Institutions –- Redundancy – System safety – Trust*

Résumé. *Cet article vise à mettre en lumière les effets de la confiance sur les performances en termes de fiabilité dans les Organisations à Haute Fiabilité (OHF) comme les centrales nucléaires, les usines chimiques ou les services d'urgence des hôpitaux. L'auteur affirme que dans l'analyse de la confiance dans le cadre de ces systèmes socio-techniques, il convient de considérer non seulement les effets bénéfiques mais aussi négatifs de la confiance en termes de fiabilité. Les effets qui résultent potentiellement d'un comportement de confiance sont discutés dans le cadre de deux types d'interaction dans la gestion des tâches dans les organisations à haute fiabilité: l'interaction humaine vs l'interaction homme–machine. La confiance est ensuite*

Social Science Information, 0539-0184; Vol. 48(2): 315–333; 102416

DOI: 10.1177/0539018409102416 http://ssi.sagepub.com

rapportée aux contraintes et aux exigences susceptibles d'interférer avec les effets bénéfiques du comportement de confiance. En particulier, l'auteur s'interroge sur trois types de croyances qui modulent l'effet de la confiance sur les performances en termes de sécurité: les croyances basées sur des valeurs et des normes communes, les croyances basées sur l'institution et les croyances basées sur la fiabilité du système. Finalement, l'auteur met en lumière les facteurs organisationnels qui se révèlent essentiels pour le développement et le maintien d'environnements de travail sûrs dans lesquels les aspects bénéfiques de la confiance peuvent s'exprimer.

Mots-clés. *Confiance – Institutions – Interaction – Organisations à haute fiabilité – Redondance – Sécurité des systèmes*

The aim of this article is to illuminate the role of trust in a specific type of organization, namely high-reliability organizations (HROs), like nuclear power plants, chemical plants or hospital emergency departments. In such organizations intrinsic hazards are always present. Multiple personnel with varied expertise work together in units and teams guided by systems, structures and procedures conducive to safety and reliability. Although the definitions of HROs in recent literature differ to a fair degree (e.g. Marais, Dulac & Levenson, 2004), I use the term HRO to generally refer to organizations that actively manage to control the risks of technical operations and depend on maintaining high levels of performance reliability and safety (Rochlin, 1993).

Literature from a diversity of disciplines highlights that trusting behavior significantly improves organizational effectiveness, for instance cooperation and communication (e.g. Dirks & Ferrin, 2001), adaptive organizational forms like network relations (Miles & Snow, 1992; McEvily, Perrone & Zaheer, 2003), resource exchange between units (Tsai & Goshal, 1998) and managerial promotion of empowerment (Schoorman, Mayer & Davis, 1996). Accordingly the concept of trust has started to emerge as an important topic in HRO research. For instance, empirical studies suggest trust to be an important mediator for the influence of leadership on safety performance (e.g. Donald & Young, 1996; Zohar, 2002; Clarke & Ward, 2006). The more leaders promote trust from their workers, by inspirational appeals for instance, the more workers reciprocate and increase their commitment to safety, by actively participating in safety campaigns and so forth. Reason (1998) discusses the role of trust in reporting systems within HROs where workers are requested to report errors and near misses for the sake of organizational learning. Specifically, the success of such feedback functions crucially depends on how workers trust they will receive fair treatment from the management and their work mates. Moreover, trust is a central component in models of safety culture, which implies that behavioral norms, shared

assumptions and values of organizational members constitute a major source of system safety (e.g. Helmreich & Merritt, 1998; INSAG-15, 1998). In contrast, low trust relations between key stakeholders are assumed to have a negative impact on safety culture by reinforcing blame and fostering non-reporting of safety-relevant information (Cox, Jones & Collinson, 2006).

On the other hand, it has recently been proposed that trust may have detrimental effects on safety performance (e.g. Jeffcott et al., 2006; Conchie & Donald, 2008). Therefore in a special issue on 'Trust in high-risk work contexts', published in *Risk analysis*, Conchie, Donald & Taylor end their editorial on 'Trust: the missing piece in the safety puzzle' with a more cautious evaluation of the concept:

> In a safety-critical work environment, it is important to promote moderate levels of trust and an element of scepticism and healthy wariness. These attitudes should be promoted toward all groups, and in particular supervisors and leaders. To achieve this, programs might focus on improving beliefs and feelings about another's trustworthiness (Conchie, Donald & Taylor, 2006: 1103).

The authors provide a moderate view stressing the potential of high levels of trust to reduce personal responsibilities for safety and create an over-reliance on other people (see also Conchie & Donald, 2008). In line with this, a comparison of train-operating companies in the UK by Jeffcott et al. (2006) shows that rule-based trust at the operational level, understood as over-reliance on formal procedures, can reduce alertness to lapses in the safety system. Other scholars who explicitly address the importance of creative mistrust (e.g. Hale, 2000) and distrust (e.g. Burns, Mearns & McGeorge, 2006; Conchie & Donald, 2008) propose that trusting behavior might hinder workforces in challenging unsafe acts and lead to groupthink situations. Moreover, a consulting group of the International Atomic Energy Agency highlights the importance of a questioning attitude as one main individual contribution to a strong safety culture (INSAG-4, 1991), in other words, individuals question both their own behavior and safety responsibilities, and the behavior and responsibilities of others.

One can conclude from this that trusting behavior is not in general beneficial for safe performance in HROs. Instead the outcomes of trust have to be evaluated with respect to the different situations and targets where either low or high levels of trust are appropriate. In the present article I address the concept of trust with regard to the specific work context of HROs. In line with recent literature, I claim that not only beneficial but also detrimental effects have to be considered in the analysis of trust in HROs. The key question is how trust supports highly reliable performance and, on the other hand, how and under what conditions trust leads to individual actors taking or running risks. The

focus is on trust within human interactions and trust within human–system interactions. I suggest that trust will contribute within both types of interactions to highly reliable performance. However, beneficial effects differ with respect to the interactional context in which trust is conferred and, consequently, with respect to the beliefs on which trust is based. To illustrate these assumptions, I focus on three distinct aspects moderating the effect of trust on safety-specific outcomes: beliefs based on shared values and norms, institution-based beliefs and beliefs based on system reliability. Note that the article does not claim to provide a full review of previous trust-safety related research. Rather, it presents three distinct routes of trust, in order to illustrate how trust beliefs calibrate the impact of trust on safe and reliable performance.

In what follows I define the key terms used throughout this paper. I then explain why the effects of trusting behavior vary according to the type of interaction involved (human interactions vs. human–system interaction), and which particular aspects of the working context can interfere with a beneficial function of trust. Finally, I draw conclusions about the potential of trust for optimizing safety in HROs.

Conceptualizing trust in HROs

In line with recent conceptualizations of organizational trust (Mayer, Davis & Schoorman, 1995; Rousseau et al., 1998), trusting behavior is defined here as the behavioral manifestation of a trustor's intention of becoming vulnerable. Besides the individual's disposition or propensity to trust, a trusting intention is based on a positive expectation that a trustee is able and willing to act in line with the trustor's interests.

In the following section, I address the potential foundations of trust involved in human interactions versus human–system interactions.

Trust in human interactions

Organizational research on trust has identified several antecedents that foster the emergence of interpersonal trust within organizations. In general, three aspects of trust are differentiated (Mayer, Davis & Schoorman, 1995; Sitkin, 1995). An individual's trust may be based on the trustee's *competence* (ability or capability) to act as expected. *Benevolence*-based trust refers to the trustee's concern and goodwill to do the best in order to fulfill action expectations. Trust can be further based on perceived value congruence between trustor and trustee. Sitkin (1995: 188) defines *value*-based trust as obtaining when 'the

other party's beliefs and values are perceived as being congruent with your own such that they will approach unforeseeable situations in a way that is consistent with the general thrust of one's expectations'. All bases of trust have the potential to directly influence a trustor's expectation and beliefs about the other's trustworthiness and willingness to engage in trusting behavior.

In reviewing research on safety and trust, Conchie & Donald emphasize the importance of benevolence- and integrity/value-based trust for the development of shared safety values and attitudes. They contrast the safety-specific function of competence-based trust against trust based on integrity and benevolence:

> For example, trust in another's technical competence creates a confidence that another person has the necessary training to complete a job safely. However, it does not indicate whether the person will carry out the job in a safe way, openly communicate about mistakes or engage in helping behaviors (Conchie & Donald, 2008: 101).

Hence both authors assume that relational forms of trust (benevolence- and integrity-based trust) have more positive effects on safety than its rational form (competence-based trust).

Trust in human–system interactions

Whereas trust based on another party's competence, benevolence and values forms trustworthiness beliefs directly related to another individual or group, a more abstract foundation of trust becomes relevant when humans interact with rules or systems: this is *institution*-based trust. Since HRO production processes are inherently hazardous, the stakes of faulty trust decisions are high. Consequently, HROs are particularly concerned with minimizing uncertainties or known risks by high levels of standardization (Grote, 2007). For instance, formal work procedures and hazard-warning systems or safety-management systems are designed to increase reliability. These rule-based systems organize collective risk-monitoring and risk-coping activities in HROs to a large extent. When considering the safety-specific function of trust, it seems important to consider not only the trustor's expectations about other people but also the expectations about the functioning of organizational safety systems (Jeffcott et al., 2006).

In general, institution- (or system-) based trust serves as a substitute for interpersonal trust (Zucker, 1986; Shapiro, 1987). It builds on the expectation that a system is functioning and that others also trust in that function (Luhmann, 1979). The risk of misplacing trust is lowered to a tolerable level by institutions (defined here simply as a set of rules), since they guide and

constrain organizational members' way of trusting in others. Moreover, institutions or regulations have the potential to engender relational trust by increasing the degree of mutual understanding (Sitkin, 1995). In their model of initial trust formation, McKnight, Cummings & Chervany (1998: 478) define institution-based trust as the belief that proper impersonal structures are in place to enable the anticipation of a successful future endeavor. They propose that institution-based trust is a separate construct from beliefs about another party's favorable attributes. However, institution-based trust has the potential to affect trusting beliefs and intention. This is mainly due to structural assurance beliefs (i.e. that proper contextual conditions such as promises, contracts, regulations are in place) as well as situational normality beliefs (i.e. that 'things occur normally'). Both types of beliefs contribute to a trustor's perception of situations as trustworthy or not (McKnight & Chervany, 2006).

Safety research has rarely focused on the effects of institution-based trust. As I said above, Jeffcott et al. (2006) refer to safety effects of extensive rule-based trust. They assume that this type of trust involves the risk of reduced flexibility of organizational members when coping with unexpected risks. Here, especially those risks that are not covered by formal rules seem to be relevant (see also Conchie & Donald, 2008).

A related approach to trust in systems has been developed in the context of psychological automation research, which focuses on trust-relevant characteristics of automated systems. Lee & See (2004) suggest that attributions of trustworthiness to an automated system stem from the direct observation of its 'behavior' (*performance:* reliability of the automation, or 'what the automation does'), from the understanding of its underlying mechanisms (*process*: algorithms and operations of the automation, or 'how the automation operates') and from its intended use (*purpose*: the goal to be achieved by automation design, or 'why the automation was developed'). Lee & See state that these beliefs parallel dimensions of interpersonal bases of trust and contribute to the development of functional or dysfunctional levels of safety-specific trust in automated systems.

In sum I propose that effects of safety-specific trust are prone to contextual aspects of HRO work situations. Trusting beliefs (about a trustee's competence, benevolence and values) as well as institution-based beliefs (about structural assurances and normality of a situation) are assumed to moderate a trustor's intention to trust, i.e. to become vulnerable. In order to further specify the role of trust in safety performance, two general assumptions are made. First, beneficial effects of trust are identified when trust supports reliable performance of human actors in HROs. Second, detrimental effects are assumed when trusting behavior leads to unreliable performance. Both assumptions center on the notion of compensating for human and technical

failures. Whereas beneficial effects of trust involve compensating for others' failures, detrimental effects result from non-compensation, such as when others' behavior is not checked or monitored (Conchie & Donald, 2008).

Three routes of safety-specific trust in HROs

Trust has the potential to influence reliable performance in HROs both positively and negatively. In this section, I address three selected mechanisms through which trust affects safe performance. First, the safety effects of trust within human interactions: in line with the theory of trust in safety settings, I argue that relational forms of interpersonal trust (based on shared values and norms) are crucial for increasing and decreasing reliable performance. Second, I discuss the safety-specific function of trust in human–system interaction where human actors interact within institutional contexts. The role of institution-based trust in promoting safety is explored. Third, I examine trust with regard to situations where the target of trust is not a human actor but rather the system itself, focusing on the calculative function of trust. Finally, I draw some conclusions concerning those organizational factors that have the power to make beneficial effects of trust on safety more likely to occur.

Safety-specific trust in human interactions

When considering the safety-specific function of trusting behavior in HROs, it is crucial to focus on the actual context in which the relationship between trustor and trustee is embedded. Trusting another person always includes a situational analysis, which allows determining the extent to which the other is trustworthy, that is, he will show the expected behavior in a given situation (Hardin, 1992). This predictive function of trust seems to be particularly relevant when considering effects of trusting behavior in HROs. These systems have no choice but to function safely because failures might result in severe consequences. When individuals trust within risky settings, they have to adapt their trusting behavior to the potential risks implied in a current work context, in other words, they compare the level of risk (or how much may be lost in a given situation) to the level of trust, giving stronger weight to the latter. According to Mayer, Davis & Schoorman (1995), the level of risk inherent in interpersonal trusting behavior is twofold. On the one hand, it is determined by the relationship with the trustee, i.e. how that person's ability, integrity and benevolence in terms of showing the expected behavior is perceived. On the other hand, it refers to risks lying outside this relationship.

A study by Kleindienst & Schöbel (2005) shows how both characteristics influence a trust decision in a safety-critical situation, in this case, a surgical operation where human actors with different professional backgrounds have to collaborate. Using an online survey, the authors confronted 152 anesthesiologists from different German hospitals with the following decision problem:

> Imagine you are working together with a surgeon. In the course of the operation you notice a loss of blood in the patient. There is no unit of cross-matched blood in the operating room, which would be necessary for an immediate transfusion. The surgeon says that he has everything under control, which implies that he is able to stop the bleeding sufficiently.

The dilemma of the protagonist can be mapped onto the following decision: to trust the surgeon, or to order a unit of cross-matched blood instead. The results indicated that participants revealed significantly more trust (no order of a unit of cross-matched blood) with a young patient with better health status compared to an older, less stable patient. Moreover the participants showed a higher degree of trust when the trustee had a higher hierarchical position than themselves, compared to a condition where trustor and trustee had the same rank. These results demonstrate that the magnitude and the probability of relational and situational risks are of crucial importance for an actor's trusting behavior in HROs.

When considering the safety-specific function of trust within this type of human interaction, two arguments can be made. The first rests on the assumption that trust behavior is beneficial for safety in that it produces and is based on shared safety perceptions and attitudes (e.g. DePasquale & Geller, 1999; Watson, Bishop & Scott, 2005; Conchie & Donald, 2008). In line with this assumption, I propose that value-based trust has the potential to positively influence safety. Trust promotes safety on a group level in that it allows for collaboration between the trustor and the trustee (Mishra, 1996). Since anesthesiologists and surgeons have different professional backgrounds, their overlapping knowledge and perspectives on the same situation merge through ongoing trusting behavior, which is a necessary requirement in the face of the uncertainty of the situation. Weick & Roberts (1993) conceptualize such highly reliable team performances as the result of 'heedful' interaction between individuals, their shared knowledge and task responsibilities. Moreover, value-based trust allows for safety-promoting collaboration by broadening the behavioral repertoire of a trustor. For instance, it includes the option to trust and to compensate (to order a unit of cross-matched blood) when the trustee's performance is assumed not to be reliable. In that sense, the anesthesiologist can feel safe in receiving criticism, or speaking up if he expects a negative outcome from the behavior of the surgeon. This intrapsychic state parallels the concept of psychological safety (Edmondson, 1999). This concept describes individuals' perceptions in teams about the consequences of interpersonal risks (e.g.

being seen by others as ignorant, incompetent or disruptive), when they seek feedback or help, give advice or discuss errors. According to Edmondson (2003), trust has positive effects on psychological safety since trusting individuals are likely to believe they will benefit from the doubt about others' safety behavior.

On the other hand, trust may have detrimental effects when trust decisions are not balanced with situational demands, in other words when unreliable performance (e.g. a human or technical failure) is not compensated by trust. For instance, Foushee (1982) illustrates dysfunctional types of hierarchy-driven interactions in multipiloted aircraft cockpits. He refers to influences on trust that do not build on shared safety values alone but also operate on a group level. In his analysis he stresses that subordinate crewmembers can become 'conditioned' not to speak up (and to trust in the competence of their hierarchically higher-ranked captain), referring to the following report where a co-pilot sums up the causes of a near-miss:

> The captain said he had misread his altimeter and thought he was 1000 ft. lower than he was. I believe the main factor involved here was my reluctance to correct the captain. This captain is very 'approachable' and I had no real reason to hold back. It's just a bad habit that I think a lot of copilots have of double-checking everything before we say anything to the captain. (Foushee, 1982: 1063)

As Foushee illustrates, cultural norms constrain trust-building and knowledge-sharing by human actors in high-reliability settings. Categorizing others in terms of their role, responsibility or expertise may lead to dysfunctional types of trust. For instance, when the trustor holds back important information in favor of information provided by the more powerful trustee**.**

It becomes evident that both beneficial and detrimental effects of trust on safety may result from group-level processes. The safety impact of trust is calibrated by shared values and rests on experiential group-specific knowledge gained in direct human interactions.

The next section discusses influences of trust on safety due to institution- and system-based beliefs inherent in human–system interactions.

Safety-specific trust in human–system interaction

HROs are conceived of as interactively complex and tight-coupled systems (Perrow, 1984). This allows greater functionality and efficiency, but produces unfamiliar and unexpected sequences of events, which are neither visible nor immediately comprehensible. In addition, single-component failures have the potential to rapidly affect several parts of the whole system simultaneously. Thus, operators in HROs are constantly required to monitor system performance

accurately, detect dysfunctional processes and, if necessary, initiate corrective actions. In HROs, monitoring of potential process deviations is formally organized by rules and systems, especially in those organizations that have a high hazard potential (e.g. nuclear power plants). Monitoring rules and systems are used to obtain real-time as well as periodical information on safety-critical processes and system states.

In order to discuss safety-specific trust within human–system interactions, I differentiate between redundant monitoring systems and automated systems. The two differ with respect to the main target of trust (trust in humans within institutional contexts vs. trust in the system itself).

Trust within redundant monitoring systems

In HROs with high levels of standardized risk control, the pooling of individuals' information-processing and distribution is organized primarily through bureaucratic systems based on the redundancy principle. The impact of individual failures (e.g. misinterpreting safety-relevant information or not checking system states) on system reliability is meant to be minimized by structuring collective risk monitoring in line, i.e. in sequence. For instance, in the context of risk-monitoring systems in nuclear industries, individuals receive information about co-workers' decisions and behavior in written and condensed form (via monitoring sheets, shift books, IT systems or safety reports). In this case to trust means to rely on others' past behavior as a source of valid information about a current system state.

Beneficial effects of trust can be directly linked to the concept of institution-based trust. High levels of institution-based trust allow the trustor not to be concerned about a trustee's benevolence and integrity, since these systems substitute for relational aspects of trust. Contrary to pure human interactions where compensating for others' failures is up to the trustor's beliefs about others' benevolence and values, here the system structurally assures compensatory behavior and therefore reduces interpersonal risks of misplaced trust because less is at stake (Sitkin, 1995).

However, high levels of institution-based trust may also have detrimental effects on safety. Two lines of explanation appear to be relevant here. First, in HROs safety is not visible to human actors, in other words safety is taken for granted when there is no deviance from expected system performance. A co-worker who does nothing but his regular work performance turns out to be a safety-relevant cue for a trustor. For instance, a co-worker confirms the installation of a new component in the IT system. High levels of institution-based trust may foster the assumption that the other has already checked the

component and did not detect an unsafe state. This assumption will be more likely the stronger the situational normality beliefs are in place. However, relying on the 'unsuspicious' behavior of others may be based on a faulty assessment. According to the phenomenon of social shirking (Sagan, 2004), individuals tend to shirk unpleasant duties because they tend to assume that someone else will naturally take care of the problem. The result is a diffusion of responsibility, stemming from a particular type of institution-based trust, that is, a belief in the regularity of others' work activities.

Second, trust in information that has been directly inferred from others' behavior within institutional contexts may lead to reliability losses. Imagine a nuclear power-plant operator assigned to monitor deviations from defined parameters. He obtains critical information signaling an unsafe system state. But this information is only probabilistically related to the true system state. As the operator knows, his diagnosis will be correct with a certain probability only. Looking at the monitoring sheet, he notices that his supervisor and his colleague have recently diagnosed a safe system state. What should the operator do? Should he rely on his own actual signal or instead trust in the information provided by his supervisors and colleagues? In reality the operator has to base his trust decision on incomplete information; he has to speculate about what kind of information the supervisor and his colleague, respectively, have based their diagnosis on. Specifically he has to consider that both predecessors may have relied on different information and that his colleague may have behaved in conformity with the hierarchical superior and/or more-knowledgeable supervisor. However, even when the supervisor and colleague have decided according to their own signal (i.e. signaling a safe system state), it is still possible that all of the actors (including the trustor) are wrong. As the theory of information cascades (Bikhchandani, Hirshleifer & Welsh, 1992; Anderson & Holt 1997) holds, initially misrepresentative information may start chains of incorrect decisions which are not going to be interrupted by more representative information gained later. By means of Bayesian modeling of sequential decision-making processes, Schöbel & Rieskamp (under review) were able to show that both normative influences (through hierarchical effects) and informational influences (through expertise) can enhance the decision weight of trusted social information, which leads to a higher probability of faulty decision-making patterns. Given all these pitfalls, the crucial point is that a decision to trust within an institutional context always implies uncertainty about the trusting behavior of others. Although redundant control systems have the potential to minimize uncertainties and to substitute for and support interpersonal trust, they can allow and sometimes even foster uncertainties, in particular when hidden social influences lead the trustor to underestimate potential

uncertainty in the trustee's action. As a consequence, collective risk-monitoring performance may decrease due to redundant communication structures that involve a high level of institution-based trust.

Trust in automated systems

In the area of human interaction with automated systems, the notion of trust is of prime importance. While collaborative interactions involve continuous and mutually adaptive processes between human actors, trust in automated systems is primarily a one-sided affair: the target of trust is the system and no direct reciprocity between trustor and the trustee is to be expected. Therefore beneficial and detrimental effects of trust on safety rest mainly on its calculative function: the perceived reliability of an automated system. Trust can be understood here as an attitude which influences the quality and outcomes of man–machine interaction (Lee & See, 2004). From a normative point of view, the degree of trust in automation should correspond with performance features of the automated system: its reliability, transparency and usability. Mismatches between trust in the system and actual system performance result in inappropriate human monitoring and information-sampling behavior.

According to Manzey & Bahner (2005), two kinds of mismatches (i.e. detrimental effects on safety) can be identified. On the one hand, individuals tend uncritically to count on the reliability of automated systems and neglect to monitor and check system performance. This phenomenon is called complacency. On a motivational level, complacency reflects a relatively low level of suspicion toward system performance (Wiener, 1981). Consequences of complacency may include a loss of situational awareness (Endsley, 1995) and the risk that human actors fail to detect and manage automation failures in due time (Bahner & Manzey, 2008). On the other hand, individuals may in general show only low trust in automated systems, in other words, have a general tendency to undervalue the benefits of automation and instead rely on their own skills and competencies. This behavior can have severe consequences, particularly where automated warning and alarm systems are concerned. For instance, Parasuraman, Hancock & Olofinboba (1997) showed that the degree of trust in automated systems is moderated by the threshold level of alarm systems and the base-rate frequency of unsafe system states. Both variables refer to the reliability of a system and affect the perceived trustworthiness of alarm systems.

Notably, interpersonal trust can be further assumed to affect the prioritization of alarms and warnings from automated systems in HROs. This appears to be especially the case when there is low trust in the performance reliability of an automated system. Empirical evidence for this assumption comes

from a field study conducted in an Eastern European nuclear power plant by Ignatov, Wilpert & Schöbel (2001). By means of group discussions and qualitative interviews with plant personnel, the authors were able to derive systematic descriptions of safety-critical work situations in which compliance with safety rules was in question. These situations were then applied in a questionnaire in order to measure the relative importance of individual attitude, subjective norm and perceived behavioral control in the prediction of compliance with safety rules (according to the theory of planned behavior: Ajzen & Madden, 1986; Ajzen, 1991). One situation describes the handling of a potential false alarm:

> Seismic measurement channel No. 2 of one of the emergency reactor protection sets, located in the containment, generates a warning signal. The inspection of other seismic measurement channels does not show signs of real seismic hazard. Measurement channel No. 2 periodically (approximately every 30 minutes) generates warning signals and automatically switches off in 10 seconds. According to the procedures, you are obliged to perform a renewed inspection in order to make sure once again that the system is functioning properly.

The results of the study showed that both subjective norm and individual attitude are significant predictors of the behavioral intention to omit a renewed inspection. Interestingly, multiple regression analyses revealed that subjective norm makes a significantly greater contribution to the prediction of intention than individual attitude. The findings show that interpersonal trust may interfere with the interpretation of dynamic hazard warnings in HROs in situations where two distinct informational sources (i.e. warning signals vs. perceived other persons' behavior) conflict. Further evidence for this notion comes from the analysis of the mid-air collision of a Tupolev 154m and a Boeing 757 cargo aircraft in Germany in 2002 (Bennet, 2004). At the time of collision, these aircrafts were directed by the Swiss Air Traffic Control (ATC). Both airplanes were equipped with an on-board anti-collision device known as Traffic Alert and Collision Avoidance System Version 2 (TCAS II). If TCAS II senses that two aircrafts are on a collision course, it sends reciprocal instructions to both of the crews. Instruction effectiveness depends on both crews obeying the instructions or not. While the Boeing 757 crew obeyed the TCAS instruction, the Tupolev crew decided to follow the – contradictory – instruction of a human air-traffic controller, and therefore flew into the path of the Boeing aircraft. The analysis of the collision identified several important factors that contributed to the crew's decision not to trust TCAS II, for instance cultural and procedural factors concerning the introduction of TCAS II in the EU and Russia. Bennet's analysis further shows that conflict between trust in others and trust in automated systems may lead to dramatic consequences.

In sum, it becomes evident that trusting behavior in HROs is also manifested in human–machine interaction. The beneficial effects of trust are shown to depend on the correspondence between the degree of trust and the performance characteristics of automated systems. Trust mediates human actors' alertness in order to compensate for technical hazards. However, due to the opacity of automated systems, trust mismatches may occur which undermine the benefits of automation. In contrast to interpersonal trust, where trustor and trustee are to a certain degree aware of each other's intentions and behavior, such symmetry does not apply to man–machine interactions (Lee & See, 2004). Contrasting the available information about the trustworthiness of a human with that of a technical trustee, it seems to be more plausible to trust the human actor than the technical actor due to attributional variance and associated uncertainties inherent in automated systems.

Conclusion

The article has tried to specify relevant facets of trust that are manifested in human interactions and human–system interactions in HROs. It highlights the idea that beneficial and detrimental effects of trust on safety performance vary according to the context in which trust is conferred. In general, trust emerges as a crucial component of safety performance in HROs. Accordingly three potential routes are explored through which trust affects safety performance in HRO settings.

First, when human actors interact, both beneficial and detrimental effects of trust can be attributed to group-level processes based on shared values and norms. The potential of value-based trust is to foster collaboration between 'psychologically safe' actors. In contrast, detrimental effects may result when trust relations are socially forced by values that undermine safety. It thus becomes evident that shared values are important safety calibrators of trusting behavior in HROs. This assumption is based on models of safety culture that highlight the importance of trust within an effective safety culture (e.g. Reason, 1998; Cox, Jones & Collinson, 2006). For instance, Watson, Bishop & Scott (2005) showed that shared employee norms as well as trust in supervisors and beliefs in management safety values are important predictors of workplace safety. The authors assume that these predictors constitute the relational dimension of social capital (Coleman, 1988; Nahapiet & Goshal, 1998) necessary to promote workplace safety.

The fact that safety-promoting collaboration is facilitated by trust is also reflected in the concept of requisite variety (Ashby, 1958; Weick, 1987). Based on the assumption that humans are not as complex as the systems they

have to manage, Weick (1987) argued that 'when people have less variety than is requisite to cope with the system, they miss important information, their diagnoses are incomplete, and their remedies are short-sighted and can magnify rather than reduce a problem' (p. 112). In order to overcome the mismatch between (less complex) individuals and (highly complex) systems, he suggests pooling differing individual observations on system states by means of social interactions and networks. According to Weick, trust has the power to enhance collective requisite variety because it enlarges the pool of informational input.

As a second important aspect of safety-specific trust, institution-based trust is considered. On the one hand, institutional contexts support interpersonal trust by reducing interpersonal risks, especially those that become salient when failures of other human actors are detected. Formalized monitoring systems based on the redundancy principle allow for compensating for others' failures since they structurally assure that it is appropriate to compensate for human and technical failures. On the other hand, institution-based beliefs have the potential to bias trust beliefs about others' past behavior. Specifically, risk-monitoring systems can provide only incomplete information about the other's trustworthiness. Strong beliefs in the normality (i.e. safety) of a trusting situation may lead to misinterpretations of behavior, for instance, trusting a co-worker's signature as evidence of performed checks without knowing whether the check was really performed. Overt behavior and inferred judgments of co-workers become relevant informational cues, which are taken as highly trustworthy, irrespective of potential uncertainties involved in the behavior of the trustee. In addition, informational as well normative social influences are hidden by bureaucratically organized risk-monitoring systems. These risks should be especially likely to occur in ultra-safe systems like nuclear power or chemical plants, where major accidents are rare events (Amalberti, 2002).

When considering the safety specifics of institution-based trust, it is important to note that the organizational members' beliefs about the function of a safety system should also correspond with their shared values in order to guarantee the system's expected benefits. For instance, the success of reporting systems in HROs crucially depends on how blame and punishment are handled (e.g. Reason, 1998; Cox, Jones & Collinson, 2006). Perceived unfairness, lack of commitment on the part of top management or assigning blame to the system user are system outcomes that may counteract the intended function of a system and, therefore, lower structural assurance beliefs. Sitkin (1995) proposed that regulations can undermine the opportunity to gain trust-related benefits by goal-displacement, that is to say 'viewing legalistic procedures as the end rather than a means to fostering high

levels of relational trust' (p. 207). It becomes evident that beneficial effects of trust in institutional HRO contexts are prone to the shared values of organizational members, irrespective of their inherent trust-relevant characteristics. In contrast to other types of organizations, HROs have the difficult task of aligning and balancing two sets of member values: values supporting efficient production processes and values supporting safety, which sometimes are at stake with production concerns. Therefore the promotion of beneficial trust in HROs presupposes carefully disentangling both aspects when dealing with human interactions and human–system interactions.

Last but not least, safety-specific trust refers to human interactions with automated systems. Beneficial and detrimental effects of trust rest mainly on the trustor's capability to accurately assess the reliability of an automated system. Due to the opaque properties of technical actors, individuals can be expected to show optimal trust in automated systems as long as the functioning of these systems corresponds to the mental models of their users. In line with this, a new generation of automated systems has been recently claimed, to which the role of a 'team-partner' is assigned such that human actors can interact with them in a collaborative way (e.g. Christoffersen & Woods, 2002).

All in all, trust emerges as a most valuable and still highly underrated concept for the optimization of safety performance in HROs. More specifically it becomes evident that we should abstain from conceiving of trust as a general, context-independent remedy. Instead the full benefits of trusting behavior can be gained only if HROs manage to develop and maintain work settings that favor the beneficial aspects of trust, including collaboration and knowledge-sharing between human actors in HROs.

Markus Schöbel is Scientific Assistant in the Department of Work, Engineering and Organizational Psychology at the Berlin Institute of Technology. His current research interests include safety culture, safety management and sequential decision-making processes in high-reliability organizations. *Author's address*: Berlin Institute of Technology, Department of Psychology and Ergonomics, F7 Marchstrasse 12, 10587 Berlin, Germany. [*email*: markus.schoebel@tu-berlin.de]

References

Ajzen, I. (1991) 'The theory of planned behavior', *Organizational behavior and human decision processes* 50: 179–211.

Ajzen, I. & Madden, T. J. (1986) 'Prediction of goal-directed behavior: attitudes, intentions, and perceived behavioral control', *Journal of experimental social psychology* 22: 453–74.

Amalberti, R. (2002) 'Revisiting safety and human factors paradigms to meet the safety challenges of ultra complex and safe systems', in B. Wilpert & B. Fahlbruch (eds) *System safety: challenges & pitfalls of intervention*, pp. 265–76. Amsterdam: Elsevier Science.

Anderson, L. R. & Holt, C. A. (1997) 'Information cascades in the laboratory', *The American economic review* 87(5): 847–62.

Ashby, W. R. (1958) 'Requisite variety and its implications for the control of complex systems', *Cybernetica* 1(2): 83–99.

Bahner, E. & Manzey, D. (2008) 'Misuse of automated decision aids: complacency, automation bias and the impact of training experience', *International journal of human–computer studies* 66: 688–99.

Bennet, S. (2004) 'The 1st July 2002 mid-air collision over Überlingen, Germany: a holistic analysis', *Risk management* 6(1): 31–49.

Bikhchandani, S., Hirshleifer, D. & Welch, I. (1992) 'A theory of fads, fashion, custom, and cultural change as informational cascades', *Journal of political economy* 100: 992–1026.

Burns, C., Mearns, K. & McGeorge, P. (2006) 'Explicit and implicit trust within safety culture', *Risk analysis* 26(5): 1139–50.

Christoffersen, K. & Woods, D. D. (2002) 'How to make automated systems team players', *Advances in human performance and cognitive engineering research* 2: 1–12.

Clarke, S. & Ward, K. (2006) 'The role of leader influence tactics and safety climate in engaging employees' safety participation', *Risk analysis* 26(5): 1175–85.

Conchie, S. M. & Donald, I. (2008) 'The functions and development of safety-specific trust and distrust', *Safety science* 46: 92–103.

Conchie, S. M., Donald, I. & Taylor, P. J. (2006) 'Trust: missing piece(s) in the safety puzzle', *Risk analysis* 26(5): 1097–104.

Coleman, J. S. (1988) 'Social capital in the creation of human capital', *The American journal of sociology* 94: 95–120. (Supplement: 'Organizations and institutions: sociological and economic approaches to the analysis of social structure')

Cox, S., Jones, B. & Collinson, D. (2006) 'Trust relations in high-reliability organizations', *Risk analysis* 26(5): 1123–38.

DePasquale, J. P. & Geller, S. E. (1999) 'Critical success factors for behavior-based safety: a study of twenty industry-wide applications', *Journal of safety research* 30(4): 237–49.

Dirks, K. T. & Ferrin, D. L. (2001) 'The role of trust in organizational settings', *Organization science* 12(4): 450–67.

Donald, I. & Young, S. (1996) 'Managing safety: an attitudinal-based approach to improving safety in organizations', *Leadership and organizational development journal* 17: 13–20.

Edmondson, A. C. (1999) 'Psychological safety and learning behavior in work teams', *Administrative science quarterly* 44(4): 350–83.

Edmondson, A. C. (2003) 'Managing the risk of learning: psychological safety in work teams', in M. West, D. Tjosvold & Ken G. Smith (eds) *International handbook of organizational teamwork and cooperative working*, pp. 255–75. Chichester: John Wiley & Sons.

Endsley, M. (1995) 'Measurement of situation awareness in dynamic systems', *Human factors* 37(1): 65–84.

Foushee, H. C. (1982) 'The role of communications, socio-psychological, and personality factors in the maintenance of crew coordination', *Aviation, space, and environmental medicine* 53: 1062–6.

Grote, G. (2007) 'Understanding and assessing safety culture through the lens of organizational management of uncertainty', *Safety science* 45(6): 637–52.

Hale, A. (2000) 'Culture's confusions' (Editorial), *Safety science* 34: 1–14.

Hardin, R. (1992) 'The street-level epistemology of trust', *Politics & society* 21(4): 505–29.

Helmreich, R. & Merritt, A. (1998) *Culture at work in aviation and medicine.* Aldershot: Ashgate.

Ignatov, M., Wilpert, B. & Schöbel, M. (2001) 'Implicit norms as regulators of safety performance', Final report, No. 1501082, BMWI–Federal Ministry of Economics and Technology, University of Technology, Berlin.

INSAG-4 (1991) *Safety culture*. Vienna: International Atomic Energy Agency (Safety series, no. 75-INSAG-4).

INSAG-15 (1998) *Key practical issues in strengthening safety culture*. Vienna: International Atomic Energy Agency (Safety series, no. 75-INSAG-15).

Jeffcott, S., Weyman, A., Pidgeon, N. F. & Walls, J. (2006) 'Risk, trust, and safety culture in UK train operating companies', *Risk analysis* 26(5): 1105–21.

Kleindienst, C. & Schöbel, M. (2005) 'Working together in the operating theatre: determinants of trust-based decisions', in C. Korunka & P. Hoffmann (eds) *Change and quality in human service work*, pp. 289–97. Munich: Hampp Publishers (Organizational psychology and health care, vol. 4).

La Porte, T. R. & Consolini, P. M. (1991) 'Working in practice but not in theory: theoretical challenges of high-reliability organizations', *Journal of public administration research and theory* 1: 19–47.

Lee, J. D. & See, K. A. (2004) 'Trust in automation: designing for appropriate reliance', *Human factors* 46(1): 50–80.

Luhmann, N. (1979) *Trust and power*. Chichester: Wiley.

Manzey, D. & Bahner, J. E. (2005) 'Vertrauen in Automation als Aspekt der Verlässlichkeit von Mensch–Maschine-Systemen', in K. Karrer, B. Gauss & C. Steffens (eds) *Beiträge zur Mensch–Maschine-Systemtechnik aus Forschung und Praxis: Festschrift für Klaus-Peter Timpe*, pp. 93–109. Düsseldorf: Symposion.

Marais, K., Dulac, N. & Levenson, N. (2004) 'Beyond normal accidents and high reliability organizations: the need for an alternative approach to safety in complex systems', Paper presented at the Engineering Systems Symposium, Massachusetts Institute of Technology, Cambridge, MA.

Mayer, R. C., Davis, J. H. & Schoorman, F. D. (1995) 'An integrative model of organizational trust', *The Academy of management review* 20(3): 709–34.

McEvily, B., Perrone, V. & Zaheer, A. (2003) 'Trust as an organizing principle', *Organization science* 14(1): 91–103.

McKnight, H. & Chervany, N. (2006) 'Reflections on an initial trust-building model', in R. Bachmann & A. Zaheer (eds) *Handbook of trust research*, pp. 29–51. Cheltenham: Edward Elgar.

McKnight, H., Cummings, H. L. & Chervany, N. (1998) 'Initial trust formation in new organizational relationships', *The Academy of management review* 23(3): 473–90.

Miles, R. E. & Snow, C. C. (1992) 'Causes of failure in network organizations', *California management review* 34: 53–72.

Mishra, A. K. (1996) 'Organizational responses to crisis: the centrality of trust', in R. M. Kramer & T. R. Tyler (eds) *Trust in organizations: frontiers of theory and research*, pp. 261–87. Thousand Oaks, CA: Sage.

Nahapiet, J. & Ghoshal, S. (1998) 'Social capital, intellectual capital, and the organizational advantage', *Academy of management review* 23(2): 242–66.

Parasuraman, R., Hancock, P. A. & Olofinboba, O. (1997) 'Alarm effectiveness in driver-centered collision-warning systems', *Ergonomics* 39: 390–9.

Perrow, C. (1984) *Normal accidents: living with high-risk technologies*. New York: Basic Books.

Reason, J. T. (1998) 'Achieving a safe culture: theory and practice', *Work and stress* 12(3): 293–306.

Rochlin, G. I. (1993) 'Defining "high reliability" organizations in practice: a taxonomic prologue', in K. H. Roberts (ed.) *New challenges to understanding organizations*, pp. 11–32. New York: Macmillan.

Rousseau, D. M., Sitkin, S. B., Burt, S. & Camerer, C. (1998) 'Not so different after all: a cross-discipline view of trust', *The Academy of management review* 23: 393–404.

Sagan, S. (2004) 'The problem of redundancy problem: why more nuclear security forces may produce less nuclear security', *Risk analysis* 24(4): 935–46.

Schöbel, M. & Rieskamp, J. (under review) 'Social influences in sequential decision-making processes'.

Schoorman, F. D., Mayer, R. C. & Davis, J. H. (1996) 'Empowerment in veterinary clinics: the role of trust in delegation', Paper presented at annual meeting of the Society for Industrial and Organizational Psychology, San Diego.

Shapiro, S. P. (1987) 'The social control of impersonal trust', *American journal of sociology* 93: 623–58.

Sitkin, S. (1995) 'On the positive effects of legalization on trust', *Research on negotiation in organizations* 5: 185–217.

Tsai, W. & Goshal, S. (1998) 'Social capital and value creation: the role of intra-firm networks', *The Academy of management journal* 41: 464–76.

Watson, G., Bishop, J. & Scott, D. (2005) 'Interpersonal dimensions of safety in the steel industry', *Journal of business and psychology* 19(3): 303–18.

Weick, K. E. (1987) 'Organizational culture as a source of high reliability', *California management review* 29(2): 112–26.

Weick, K. E. & Roberts, K. H. (1993) 'Collective mind in organizations: heedful interrelating on flight decks', *Administrative science quarterly* 38: 357–81.

Wiener, E. L. (1981) 'Complacency: is the term useful for air safety?', Paper presented at the proceedings of the 26th Corporate Aviation Safety Seminar, Denver, CO.

Zohar, D. (2002) 'The effects of leadership dimensions, safety climate, and assigned priorities on minor injuries in work groups', *Journal of organizational behavior* 23(1): 75–92.

Zucker, L. (1986) 'Production of trust: institutional sources of economic structure, 1840–1920', in B. M. Staw & L. Cummings (eds) *Research in organizational behavior*, vol. 8, pp. 53–111. Greenwich, CT: JAI Press.