This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TTS.2021.3118355, IEEE Transactions on Technology and Society

Delegation of moral tasks to automated agents – The impact of risk and context on trusting a machine to perform a task

Gian Luca Liehner
Human-Computer Interaction Center
RWTH Aachen University
Aachen, Germany
ORCID: 0000-0001-5065-1619

Philipp Brauner
Human-Computer Interaction Center
RWTH Aachen University
Aachen, Germany
ORCID: 0000-0003-2837-5181

Anne Kathrin Schaar
Human-Computer Interaction Center
RWTH Aachen University
Aachen, Germany
ORCID: 0000-0002-8643-6213

Martina Ziefle
Human-Computer Interaction Center
RWTH Aachen University
Aachen, Germany
ORCID: 0000-0002-6105-4729

***Abstract*--The rapid development of automation has led to machines increasingly taking over tasks previously reserved for human operators, especially those involving high-risk settings and moral decision making. To best benefit from the advantages of automation, these systems must be integrated into work environments, and into society as a whole. Successful integration requires understanding how users gain acceptance of technology by learning to trust in its reliability. It is thus essential to examine factors that influence the integration, acceptance, and use of automated technologies. As such, this study investigated the conditions under which human operators were willing to relinquish control, and delegate tasks to automated agents by examining *risk* and *context* factors experimentally. In a decision task, participants (N=43, 27 female) were placed in different situations in which they could choose to delegate a task to an automated agent or manual execution. The results of our experiment indicated that both, context and risk, significantly influenced people's decisions. While it was unsurprising that the reliability of an automated agent seemed to strongly influence trust in automation, the different types of decision support systems did not appear to impact participant compliance. Our findings suggest that contextual factors should be considered when designing automated systems that navigate moral norms and individual preferences.**

***Index Terms*--automation, trust, decision-making, moral decisions, risk perception, context, moral machines, machine ethics, decision support systems, artificial intelligence**

I. Introduction

Whether we like it or not, our daily lives are filled with countless interactions with automated machines. From using an automatic teller, to online purchases, to (semi-)automated vehicles, automation has pushed its way into every domain of society. It is regularly encountered in technological sectors such as robotics, aviation, medicine, automated driving, manufacturing processes, shipping, and process control. The primary aim of automation is to substitute human labour with automated processes and procedures [1], [2]. Repetitive tasks that lead to fatigue, unreliability, and inefficiency in people are replaced by machines that can perform these same tasks accurately and tirelessly [3] (as well as reduce labour costs for employers). In addition, tasks that are difficult or dangerous to perform [4] such as disarming bombs or saving people from burning buildings, can be completed without increasing risk to human life. While machines in the past were used to perform repetitive tasks on lower logical levels [5] (e.g., kitchen appliances in the 1960s), with the fast development of AI and increases in computing power, automation is becoming more common for more complex tasks. In other words, tasks once reserved for human operators because of their high cognitive demands are now routinely performed by machines. The most anticipated, and yet most controversially example is *automated driving*. The controversy is specifically linked to the uncertainty when driving, and the inevitable split-second decision-making required given the complexity of navigating real environments. Although computing power has increased, and AI is becoming increasingly more advanced, automated systems are not yet able to perform error-free tasks. In fact, it remains an open question if they ever will. The operation of self-driving vehicles in public spaces necessarily involves human lives. This means that ethical questions inevitably arise. What happens if an automated vehicle crashes into a pedestrian? Are machines able to perform such delicate and deeply moral tasks? And if so, what set of rules and beliefs are they subject to?

In recent years, substantial research has been dedicated to investigating decisions made by automated systems in situations of moral dilemma and to the social dilemmas resulting from those decisions [6]–[9]. The *Trolley Problem* is a thought experiment whose most common version was proposed by the English philosopher Philippa Foot [10]. It has been adapted and used repeatedly to investigate the social acceptance of moral decisions made by automated vehicles. Briefly, a runaway trolley is heading towards five people that are unable to move. A person (A) stands next to a lever that can deviate the train to a different track where only one person, also unable to move, is standing. Should A pull the lever, thus saving five people but killing one actively, or should A do nothing and kill five through inaction, and avoid actively participate in the killing of one person? Another prominent study is the *Moral Machines* experiment. Here, Awad et al. investigate ethical rules that self-driving vehicles could follow in situations of moral dilemma [11]. Following the trolley dilemma, subjects had to choose from possible accident outcomes in a choice experiment. They ultimately find a social preference to spare more lives over fewer, young lives over older, and humans over animals [11]. However, Bonnefon et. al [12] found even though people tended to favour self-driving vehicles programmed to save more lives, when it came to actually buy such vehicles, they were rather reluctant to do so. While large amounts of research have investigated these specific issues, little attention has been paid to the process of delegating tasks with moral values to automation, or to the decision-making processes involved in general.

For automation to be beneficial and successfully integrated into society it must first be used and accepted by humans. *We can develop as much as we want, what good does it do if nobody uses it?* According to this line of thought, it is essential to investigate the factors promoting the use of given technologies, and in this case, the delegation of (moral) tasks to automated machines. Research on the topic suggests that trust plays an essential role in the acceptance, adoption, and use of automated systems [2], [13]–[18]. In fact, trust can act as a mediator in risk situations involving the use of automation [18], especially when there is a potential for injury. That said, trust is a multi-layered construct that cannot be observed directly, thus making it difficult to measure. What can be more directly investigated are factors that influence trust itself. A meta-analysis by Hancock et. al [19] revealed three main categories of trust factors. First are *system factors* that are directly related to the performance and functional characteristics of the automated agent. Second are *human factors* related to the personality traits, personal beliefs, and values of the human operator who is relinquishing control to the automation. Third are *environmental factors* characterized by the type of executed task and the physical environment in which the task is performed. Both system factors and human factors have been extensively explored [19]–[21]. Environmental factors, on the other hand, have received little attention, and merit further exploration.

In this study, we focused on environmental factors that could potentially influence the use of automation. First, assuming that automation is moving into more high-risk environments, we studied *context* as a factor. Here, context is defined as the physical environment in which automation can be used. Specifically, we focused on automated driving in three different settings: urban, health care (hospital), and production logistics (warehouse). All contexts have acquired prominent states in research. Also, automated driving is well known by laypeople and thus most relatable. Secondly, as automation is—at least in many cases—not error-free to this date, we study the degree of *risk* involved.

To accurately measure the effect of risk and context, we developed a decision task in which respondents were presented with differently shaped situations and asked to either opt for a manual execution of the given task or to delegate the task to an automated agent. We included *decision support systems* and information about the reliability of the automated agent to help participants make their decision.

The structure of this article is as follows. After this introduction, Section II provides the theoretical background on trust, decision-making under risk, as well as moral decision-making, and decision support systems. Section III lays out the research objectives and hypotheses. Section IV presents the methodological approach, the development of the decision task, and the study design. Section V presents the results of the experiment and section VI discusses their implications. Lastly, section VII covers this study's limitations and section VIII provides a conclusion and research outlook.

II. Theoretical Background

This section first defines the construct of trust, outlines factors influencing trust, and how trust relates to the use of automated systems. The second part takes a closer look at the perception of risk and decision-making under risk followed by a short section on moral decision-making and the ethics of risk. The last section covers decision support systems, and the amount of trust people are willing to place in those systems.

A. The multi-layered concept of trust

The construct *trust* has been discussed by a multitude of disciplines (e.g., philosophy, psychology, sociology) making it difficult to pin down a general all-encompassing definition. That said, two elements are of essential importance: a *trustee* that transmits information and a *trustor* that acts upon the received information [19]. The conveyed information forms the basis of trust upon which the trustee and trustor can act [18]. Within the context of automation, Lee & See define trust as *'the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability'* [18, p. 51]. Rempel et. al [22] argue that, particularly in situations of uncertainty, the trustor will challenge the trustee to ascertain the amount of trust that can be given. If the trustee fails to satisfy the expectations of the trustor, trust can decrease. Should the trustee, on the other hand, satisfy the expectations of the trustor, then trust could either be solidified or increase. This process is commonly known as the *calibration of trust* [23]–[25]. Within the field of automation, this phenomenon refers to the correspondence between the operator's trust in automation and the actual capabilities of the automated agent [18], [23], [24]. As an

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TTS.2021.3118355, IEEE Transactions on Technology and Society

illustration, take the example of self-driving vehicles (SDVs). To date, SDVs can reach automation stages ranging from 3 to 4 as set by the Society of Automotive Engineers [26] (ranging from *level 0:* no automation, humans perform all driving tasks, to *level 5:* full automation, no human interaction needed). In the ideal user situation, the user's beliefs about the automation's capabilities will be in perfect alignment with each other. Unfortunately, misalignments are quite possible, if not likely. For example, if a driver assumes the SDV has full automation capabilities (level 5) when in actuality the vehicle registers at around level 3 or 4, the level of trust will exceed the actual capabilities of the system. This situation has the potential to lead to an over-reliance on automation, which in turn can facilitate its misuse [2], a behaviour that can lead to catastrophic results, especially in high-risk settings. In February 2020, a 38-year-old man died in an accident when his self-driving car (Tesla Model X SUV) hit a barrier at 70mph (113 km/h). A subsequent investigation by the National Transport and Safety Board found that the driver did not make any attempt to slow down the car [27]. Investigators found that at the moment of the accident he was playing on his phone and undoubtedly overestimated the capabilities of his vehicle, which led to an overreliance on the technology [2], [27]. In contrast, under-reliance results from levels of trust lower than the actual capabilities of a system and leads to the underuse of automation. In this case, the technology at hand is not (properly) used, thus inhibiting its successful adoption. Consequently, the driver would assume a driving capability of level 1 or 2 when in reality it might be much higher, thus refraining from using the SDV at all.

As automation becomes more complex, human operators experience more difficulties in understanding its underlying principles and in turn tend to inaccurately predict their behaviour [28]. Here, trust plays an essential role as it helps people handle complexity and mediates the perceived uncertainty [25]. In summary, trust is a mediator between trustor and trustee in situations of uncertainty, risk, or vulnerability during which trust can build, solidify, increase, or decrease according to the outcome of the situation.

To analyse the impact of trust, and with it, the factors that influence it, trust needs to be measured accurately. However, its complexity and lack of direct observability make it challenging to measure. In fact, the development of standardized and validated scales of trust has been neglected over the last years as most measures have been developed for individual studies [29]. Currently, two different approaches measure trust in automation. The first involves *subjective measures* that record participants' perceived level of trust with regard to a specific technology [4]. These measures are generally taken multiple times during a session, and most commonly at the end of a trial or set of trials [4]. While assessing these measures is relatively simple, two issues must be stressed: First, they pose a risk of 'social desirability bias', i.e. respondents' tendency to respond to questions in a way that will be viewed favourable by others [30]. Second, although respondents may indicate trust in a given technology, their declaration may differ from their actual behaviour, thus producing a discrepancy between the subjective measure and actual behaviour. A second approach lies in *behavioural measures* of trust. While these measures are more difficult to assess, they are less subject to social desirability and self-perception biases. One example is reliance, which Lee & See argue is a direct behavioural result of trusting automation [18] and has often been used as a proxy measure for trust in automated systems xx[4]. Although trust is often assessed through behavioural measures, research on the relationship between trust and use remains inconclusive and deserves further attention.

B. *Decision-making under risk*

In the past, risk was mostly determined by natural phenomena (e.g., earthquakes, floods, hurricanes, etc.). In modern society, the technological advancements that have made that modernity possible can also be a source of risk. At its core, risk is borne out of a situation of uncertainty for which the outcome cannot be determined due to limits in available information [31], [32]. Toma & Chiti [33], in contrast, argue that risk *'refers to situations in which probabilities can be identified for possible results'* (p. 976). This definition implies that risk can be quantified. Hence, decision-makers confronted with a situation of uncertainty can calculate the probability of a potential outcome using the available information, at least to a certain extent. From this probability, they weigh their options and opt for a specific course of action. This process is usually referred to as *decision-making under risk* [31], [34].

Two assumptions must be noted. First, the decision-maker must be provided with valid information with regard to the probabilities of the consequences of the situations at hand, as well as all possible courses of action [35]. Put differently, sending a decision-maker into a situation of uncertainty without providing either a particular resolution option or valid information about the situation or its consequences cannot be characterized as decision-making under risk. Second, risk for decision-makers can only occur if the situation is in some way linked to the choice taken by the decision-makers [32]. That is, the outcome of the situation and thus, the consequences are directly linked to the course of action taken by the decision-maker. In summary, risk requires a situation of uncertainty that is measurable by a decision-maker whose choice is directly connected to the outcome of the situation (see Fig. 1).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TTS.2021.3118355, IEEE Transactions on Technology and Society

Fig. 1: Components of risk: uncertainty, Components of risk: uncertainty, measurability, and decision. Adapted from [31, p. 36].

Decision-makers arguably have two methods at their disposal to estimate the amount of risk for a particular situation. First, they can make use of *objective risk assessment*, which is based on historical data, statistics, and the calculation of probabilities [33]. For instance, the *expected value* (EV) of a course of action can be computed by the product of the probability of an event and the consequences of the outcome [35]. Returning to our example of the self-driving vehicle and the trolley problem (see section 1 & 2a). Suppose a self-driving vehicle, due to external circumstances, has an 80% probability of hitting a tree killing the passenger onboard (Option A) and a 20% chance of crashing into a crowd of five people (Option B). If, and only if, damage was based on the number of human lives put in danger, the EV for option B would, for instance, be calculated by multiplying the probability (20%) by the number of potential casualties (5), resulting in an EV of 1. Besides the EV, the expected utility (EU) can be calculated as an additional measure of objective risk assessment. Where the EV only takes into account the direct consequences of the situation, the EU paints a bigger picture and includes other factors such as the long-term consequences of the decision [35].

A second method is *subjective risk assessment*, which depends on the decision-maker's beliefs and values, character traits, feelings, desires, experience, and habits [33]. For instance, an expert with 500.000 miles of driving experience will assess situations differently than someone who has just acquired a driving license. Similarly, people that are generally more cautious could show a lower risk threshold.

While the aforementioned concepts refer to risk in general, they can easily be reinterpreted in the context of technological risk perception and risk assessment. Renn & Benighaus [34], define technological risk as *'the likelihood of physical, social, and/or financial harm/detriment/loss as a consequence of a technology aggregated over its entire lifecycle'* (p. 295) and risk perception as the *'processing of physical signals and/or information about potential hazards and risks associated with a technology and the formation of a judgment about seriousness, likelihood, and acceptability of this technology'* (p. 295). Based on the findings of risk research, Renn & Benighaus interpret central aspects for risk perception of technologies: (1) The multidimensional assessment process centred on expected damage is only one factor among many in risk perception [34], [36]. (2) The exploitation of intuitive heuristics, the perceived characteristics of risk, the risk situation, and affective associations in the development of individual, and social risk perception of technologies [37], [34]. (3) The influence of semantic images (emerging danger, insidious danger, stroke of fate, gamble, personal thrill) [38] helps individuals to classify risk thus reducing complexity [34]. (4) Psychometric variables account for a considerable proportion of variance in risk perception e.g. [39]. (5) Although there are studies that have identified universal motivational types e.g. [40], in the context of risk perceptions it could be the case that cultural and social factors play a key role in the formation of risk perception [34], [41].

In the case of potential injury to a human being, as when SDVs are employed in public spaces, an additional, non-trivial factor comes into play: *morality*.

C. *Moral decision-making and the ethics of risk*

When making moral decisions, reasoning solely on probabilities and the expected output is insufficient. Especially when human lives are at stake it is crucial to take other factors into account that are more closely related to individual or social values. Garrigan et al. [42] view moral decisions as *'a response decision about how to behave in a real or hypothetical moral dilemma (a situation with moral rules or principles attached, where a response choice is required), or it can be a judgement or evaluation about the moral acceptability of the actions, or moral character of others, including judgements of individuals, groups, or institutions'* (pp. 80-81). Accordingly, moral decisions should be made in line with the moral framework of the society or community where it is embedded. This assumption becomes particularly relevant if automated systems are involved in decision-making processes. The role of social and ethical factors, the impact of psychometric user characteristics, the respective usage context, or the technical specification of an automated system on decision-making have not yet been clarified to a satisfactory degree. Nevertheless, with regard to framing social or ethical guidelines, discussions are currently gaining momentum in the context of research projects. However, the ethical view of risk has been tackled only hesitantly in the past. For a long time, moral philosophy was limited to an evaluation of human behaviour in precisely definable situations with the use of e.g., utilitarian frameworks [43], [44]. Moral philosophy has largely left it to decision theory to address uncertainty as a factor in human decisions, and only a few theoretical works have dealt with this question [45], [46]. Nonetheless, in the future, moral valuations and the views of possible users are needed as an orientation for decisions on risk-taking and risk-imposition [47]. The role of psychometric factors has been addressed in several studies. The general finding is that humans are sceptical about machines taking over moral decision-making tasks [48], [49]. But even if these types of studies provide insight into the attitudes of the general public, they do not provide approaches for regulating automated decisions on moral issues. Gogoll et al. [49], for example, recommend not to concentrate strictly on the technical issues but to place equal emphasis on social acceptance. Above all, they recommend investing in hands-on experience as a foil to hypothetical scenarios.

D. Trust in Decision Support Systems

Decision-making can be complex and difficult, especially in risky and morally relevant situations, Therefore, attempts are often made to compensate for human frailty through decision support. In the context of increasingly complex decisions, so-called Decision Support Systems (DSSs) are introduced, which are intended to support people by presenting decision-relevant information and advice. DSSs automate the computational part of decision problems and support users by finding, aggregating, and presenting relevant information, evaluate decision alternatives, or calculate what-if scenarios [50], [51].

DSS are at the interface between the human decision-makers and the real-world, in our case the automated technical system. As outlined above, trust plays an integral role in human interaction with technical and automated systems. Within DSSs, trust can be defined as a combination of confidence and the willingness to follow the recommendation of an intelligent decision aid [52]. Prior work showed that trust in automated processes declines if the perceived competence of the automation decreases even if the actual influence on performance is limited. Also, an unreliable component decreased trust in similar components but reduced trust did not carry over to other components [53]. Error rates of automated DSS and manual interactions influence trust and operators' self-confidence. Higher manual error rates reduce self-confidence whereas higher automation errors lead to lower trust in the automation [54].

III. Objectives and Hypothesis

In lack of research on the effect of environmental factors on trust, use, and adoption of automated systems, this study explores the impact of *risk* and *context* on the delegation of moral tasks to automated agents. For this, we developed an experiment with a series of decision tasks with moral implications. Ultimately, the tasks were aimed at revealing under which conditions participants are willing to delegate a given task and whether risk and context affect their decision process. Accordingly, we propose the following hypotheses:

H1: In situations of high risk, people will be reluctant to delegate a task to automated agents.

H2: Tasks with attached moral values (e.g., potential injury to humans) will be less **likely** to be delegated to an automated agent.

H3: The context in which a task is to be executed affects the willingness to delegate that task to an automated agent.

H4: Presented system reliability influences reliance on the automated system.

H5: The type of the DSS as an additional information source influences the compliance with the support system.

IV. Methodological Procedure

A. Decision tasks with moral values attached

To validate the aforementioned hypotheses, the current study had to be conceived in a way that ensured moral values to be attached to the decisions taken by respondents as well as the inclusion of uncertainty in which participants are adequately presented with all relevant information regarding the situation at hand. A further requirement was the ability of the decision task to model real-life decision-making accurately to fully emerge the participant into the decision-process. We selected a game-based approach that represents the various situations through *microworlds* which on the one hand minimize the complexity of real-world systems, and on the other hand, preserve the essential characteristics of the task and allow experimental variations [55].

Through an iterative process (see detailed explanation in [56]) including literature reviews, paper prototyping, and a series of pretests, a final decision task was conceptualized. In its final version, participants were presented with situations in which a package was to be transported between two places. Participants were given the choice between two modes: *automatic* or *manual* delivery. This method has been repeatedly utilized throughout the research on the reliance on automation and has shown to provide robust results [53], [57]–[60]. In our study, the automated mode was presented to participants as saving on manual labour and therefore maximizing profit. However, this came with a cost as the automated collection of the package would be less safe. In fact, following the line of thought that automation to this day is not able to operate error-free, the automated delivery in this study would not always be 100% reliable and hence would potentially damage valuable goods, thus reducing profit. Even more so, the automated agent could endanger the safety of humans and potentially cause physical harm, raising the question of morality. The manual mode, on the other hand, was 100% reliable however less beneficial as personnel costs would have to be deducted from the profit.

1) Reliability

To enable participants to estimate the success probability of the automated agent (H4), following common practice [60]–[62], three reliability levels were

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TTS.2021.3118355, IEEE Transactions on Technology and Society

introduced: *low* (60%), *medium* (80%) and *high* (100%) reliability. These levels were displayed to the participants as the fraction of errors by mileage that the automated agent had covered during its operation period. From this, participants were able to estimate the probability of the automation succeeding (or failing).

2) *Context*

To study the factor context (H3), three different environments in which tasks should be executed were designed. First, following the rise of the Industrial Internet or Industry 4.0 [63], [64] and the use of interconnected automated robots in the industry, a *warehouse*, in which a forklift was tasked with the retrieval of a package from the shop floor, was graphically presented to the participants. Second, in accordance with the fast development of automated driving, a delivery vehicle is instructed to deliver a package across town in an *urban* scenario. The last scenario addressed the *medical* domain, and a hospital floor was portrayed to create a more fragile environment. Here, a health robot had the task of delivering a liver to an operating room. The test persons did not receive any further information about the contexts.

The decision tasks in each context are shown in a birds-eye view of a warehouse, a road system of a city, or a hospital (see Fig. 2). As the actual tasks are the same, the influence of context on decision making and thus possibly different moral evaluations can be studied.

3) *Risk*

The levels of risk for studying H1 and H2 were constructed following two distinct requirements. First, risk levels needed to be gradually increased to evaluate whether higher risk levels would impact the decision process of participants. Second, the level of risk required to be tainted with a certain share of moral values to assess whether the inclusion of morality would influence the reliance on the automated agent. Consequently, four levels of risk were defined: *no risk*, *property risk*, *personal risk*, and *property and personal risk*. First, property risk was defined by the potential damage to private or public property by a person that is not the owner of that property. This condition was displayed to participants by positioning valuable goods (e.g., medical equipment for the medical context) in the various environments. In contrast, personal risk was characterized by the possible injury of the body, mind, and/or emotions of another human being and was visually represented by positioning humans (e.g.., factory workers for the warehouse context) within a specific environment. In the condition property and personal risk, the two aforementioned conditions are combined. Lastly, the no risk level solely included the potential risk to the automated agent itself.

4) *Decision support system*

Finally, we included a DSS to the decision task, with the aim of recommending a specific dichotomous action to the participant (automate the given task or execute manually). To investigate if trust is rather shaped by the nature of the trustor or the reliability of the automated vehicle, the DSS either took the form of an engineer, AI, or statistical analysis software (H5).

B. *Game mechanics*

In summary, participants were provided with various visual and textual cues to assess the situation at hand (see Fig. 2). First, the context was displayed by a graphical representation of the environment which on a second layer included the various levels of risk. Second, the reliability was displayed by the fraction of errors by mileage underneath the graphical representation of context. Third, the DSS was represented visually by icons accompanied by a textual recommendation (e.g., “I think you should choose the manual way” — Artificial intelligence). Lastly, two buttons (AUTOMATE vs. MANUAL) were displayed that let participants log in their final decision. The displayed buttons were supplemented by the foreseen profit for each type of execution in case of success. The values for the automation profit were selected at random and ranged from \$2000 to \$5000. With this information in mind, the expected value for the automation profit could be calculated by multiplying the previous experience of the vehicle (reliability) by the expected automation profit. The profit for the manual option on the other hand was obtained by either adding or subtracting 10% (in an alternating manner) to the expected value of the automation profit. After selecting one of the two options, a feedback panel resolved the situation by presenting the achieved profit (or loss) and indicating whether bystanders had been harmed in any way. According to the outcome of the situation, the profit and injuries were added to the overall balance and injury counter in the upper right corner.

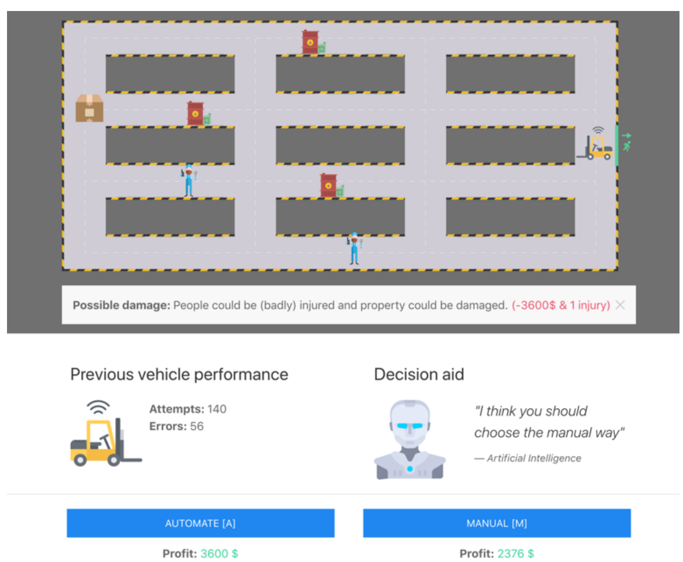


Fig. 2: Overview of the decision task illustrating the warehouse context and level of risk *property* and *personal risk*. Reliability is set to 60% and the decision support system is of type AI, recommending opting for the manual execution of the package collection by a forklift.

C. *Study design*

The study consisted of five distinct phases. In *Phase 1* demographic data and personality traits of the participants were collected. *Phase 2* provided a detailed explanation of the decision task through an explanation wizard. *Phase 3* allowed participants to undergo three training trials to accustom themselves to the decision task and the system. Training tasks were randomly chosen from the pool of generated trials. The core of the study (*phase 4)* consisted of a *3 (reliability) x 4 (risk level) x 3 (scenario) x 3 (DSS)* within-subject full-factorial design (see Table 1) resulting in a total of 108 decision tasks (trials). Based on the full-factorial design, a pool of counterbalanced trials was generated for each participant and divided into 9 subsets according to the combination of 3 (scenario) x 3 (DSS) between the sets. Thus, *reliability* and *risk level* were considered as within-set factors and *scenario* and *DSS* as between-set factors. After each set, a post-set questionnaire asked for the perception of the previous set of tasks. Lastly, *phase 5* consisted of a short semi-structured interview aimed at gaining a better understanding of the decision-making process.

TABLE I
Overview of full-factorial design

Factor	*Level*	Number of Levels
Reliability	*Low reliability (60%)* *Medium reliability (80%)* *High reliability (100%)*	3
Risk level	*No risk* *Property risk* *Personal risk* *Property & personal risk*	4
Context	*Warehouse* *Urban* *Medical*	3
DSS	*Engineer* *Statistical analysis software* *Artificial intelligence*	3

Total number of trials: 108

D. Materials & procedures

We implemented the decision tasks as a web application (detailed explanation in [56]) which allowed for a seamless integration of commonly used survey tools. Moreover, the framework provides cross-device compatibility and, due to its modular design, fast interchangeability of components as well as efficient modelling of factor combination. The final web-application was translated to German to account for German participants.

The experiment was conducted in an empty seminar room of the university on a 15-inch MacBook Pro. The web-based decision task ran on the university's server and was presented to participants through the web browser Chrome in full-screen mode. The questionnaires (pre-experiment and post-set questionnaire) were hosted by Qualtrics.

Within the decision tasks, participants were solely allowed to use the keyboard. Pressing the *A* key would select the automatic mode and pressing the *M* key would trigger the manual mode. After that, pressing the Spacebar proceeds to the next trial. We used this design to measure decision times (time between the first sight of the task and the decision of the participant) in the background relying solely on keyboard interactions to increase measurement accuracy.

E. Questionnaires

In the first questionnaire (phase 1) we asked about the participants' age, gender, current field of work (technical, non-technical domain), and their highest educational attainment. Further, we measured the *disposition to risk* queried on a validated single-item scale [65] as well as *the disposition to trust* on a validated three-item scale [66]. Complementing this measure, *disposition to trust in automation* was measured by a self-developed five-item scale based on Jian et al.'s work [67] to assess participants general level of trust in automated systems. In additions Jian et al.'s items the dimension of mistrust was added to the scale (see Appendix A).

The post-set questionnaire first asked participants how safe they felt when making decisions within the last set of trials (*safety perception*). Secondly, a risk measure was included to assess participants' *willingness to take risk* depending on the presented scenario. Lastly, participants were asked to evaluate to which extent they relied on the recommendation of the DSS (*compliance with DSS)*.

F. The sample

We recruited the participants through personal and technology-mediated social networks. They were informed that participation was voluntary, not rewarded, and that they could withdraw from the study at any time without any disadvantage (informed consent and voluntariness). In total, 43 people took part in the experiment including 27 female (62.8%) and 16 male (37.2%) participants. The average age was 25.2 years (SD = 7.5) with a minimum age of 19 and a maximum of 68 years. The sample was relatively well-educated with more than half of the participants (22, 51.2%) in possession of a high school degree, 13 (30.2%) bachelor graduates and 7 having successfully obtained their master's degree (16.3%). One person stated that they had completed an apprenticeship. Regarding their job nature, 21 participants (48.8%) claimed to work in a technical field, while 17 (39.5%) were employed in a non-technical domain. Five participants stated that they were currently unemployed. As we used a randomized full-factorial design, systematic biases through the (un-)employment status are rather unlikely. The average disposition to risk assessed in the pre-experiment questionnaire reached a mean of 5.53 (SD = 1.95) on a scale from 1 to 10, where lower scores represent higher risk aversion.

G. Analysis

Descriptive statistics are reported for the various variables. Differences in means are analysed by t-tests, one-way and two-way repeated measures analysis of variance (ANOVA) and effect sizes are reported as η^2_G. If the assumption of sphericity is not met, we report Greenhouse-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TTS.2021.3118355, IEEE Transactions on Technology and Society

Geiser and Huyn-Feldt corrected values accordingly. According to conventional practice in social science studies [68], [69], we set the level of significance for rejecting the null-hypothesis to $\alpha = .05$. To make the binary decision to automate illustrative and due to the explanatory nature of this study, the decision to automate is analysed individually for each factor using subsequent ANOVAs by aggregating the decisions of the other conditions into a metric variable ranging from 0% (no) to 100% (full reliance). We have previously assessed, using a logistic regression across all factors and the binary outcome variable, that this approach yields similar($\chi^2 = 2058$, $p < .001$, $AIC = 4355$) and have corrected for type I error inflation. We used R 4.0.3 for the analyses.

V. Results

This section presents the results of the experiment. First, we review the impact of the level of risk on the reliance on the automated agent. Second, we analysed the effect of context, i.e., the environment in which the task is embedded. For both factors (risk and context) descriptive and inferential statistics are provided and supplemented by the qualitative findings of the post-experiment interviews. Additionally, the subjective evaluation of the decision tasks collected in the post-set questionnaire is presented for the block-factor (see section 4B) context. Lastly, interaction effects are presented.

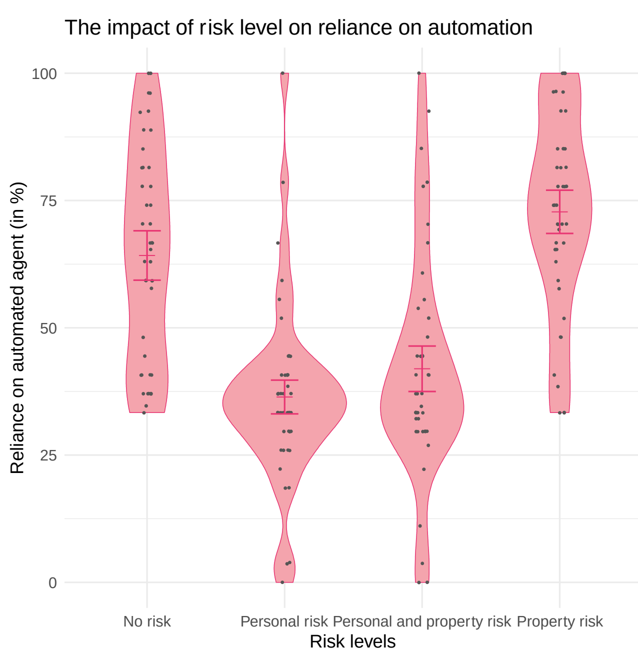
On average participants relied on the automated agent in about half of the decision tasks ($M = 53.9\%$, $SD = 14.8$). The median time to reach a decision was on average 4.29*s* ($SD = 1.97$). In the post-set questionnaire, participants were asked how safe they felt in the proposed environment (*perception of safety*), how risky they estimated their decision behaviour (*risk perception)*, and about the extent to which they relied on the previous experience of the automated agents (*reliance on reliability*). While the reliance on reliability reached a mean score of 4.58 ($SD = 1.04$), the perception of safety scored an average of 4.22 ($SD = 0.90$); both are above the centre of the scale. The mean willingness of participants engaging in risky behaviour, on the other hand, was 3.11 ($SD = 0.99$) which here is lower than the centre of the scale.

A. The impact of risk

Throughout the experiment, participants were subjected to four different gradually increasing levels of risk (H1 and H2): *no risk*, *property risk, personal risk*, and *property and personal risk*. The highest reliance on the automated agent was found in the condition in which only foreign property could be damaged (property risk) with a mean of 72.78% ($SD = 19.64$). In contrast, the condition with the lowest reliance on the automated agent was shown for the personal risk level ($M = 36.40\%$, $SD = 17.70$). Similar scores could be found for the risk level which included the potential damage to humans and foreign property with an average reliance of 41.96% ($SD = 22.28$). The level for which solely the automated agent could be damaged but without damage towards persons or property scored an average reliance of 64.21% ($SD = 21.08$).

The overall effect of risk was assessed by performing a one-way repeated measures ANOVA with the risk level as independent and the aggregated reliance score as dependent variable. The results of the analysis of variance revealed a strong and significant effect of risk on the reliance on the automated agent ($F(1.93, 81{,}13) = 51.20$, $p < .001$, $\eta^2_G = 0.88$). The results of subsequent planned orthogonal contrasts are twofold: First, the reliance on automation is significantly different in the presence of risk in comparison to situations in which risk is absent ($t(126) = 4.91$, $p < .001$). Second, the potential of damage to a human being significantly reduces the reliance on automation in comparison to the potential damage to foreign property ($t(126) = 11.26$, $p < .001$). Comparing the risk levels *personal risk* and *property and personal risk* did not yield any significant differences in reliance. Finally, analysing the effect of risk levels on decision time did not provide any evidence of significance. A comparison of average reliance for each risk level is shown in Fig. 3, which shows a violin plot indicative of the reliance distribution for each risk level.

These findings are further supported by the post-experiment interviews in which participants stated that there was not any kind of monetary value that could justify the injury to a human being. A handful of participants further stated that the potential injury of bystanders would make their decision more difficult and, in that case, would rather opt for a manual execution of the task. Moreover, it was noted that a differentiation between the personal risk and personal and property risk levels was not made, as the inclusion of potential harm to people would overshadow the damage to foreign property. Lastly, participants claimed that property damage was neglectable and that they would often act in a riskier manner in that particular case.



This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TTS.2021.3118355, IEEE Transactions on Technology and Society

Fig. 3: The impact of risk on the reliance on the automated agent including error bars and mean in violin plot.

B. *The influence of context*

To study hypothesis H3, the physical environment of the decision task was altered, and three different scenarios were presented to the participants: *warehouse*, *urban*, and *medical* scenario. With 55.86% (SD = 15.40), the highest reliance was found in the warehouse scenario in which a forklift was tasked with collecting a package on the other side of a warehouse. In the medical scenario, the average reliance was the lowest with 51.87% (SD = 16.29). Here, the medical context was represented by a hospital in which a robot was given the task of transporting a liver to the OR, making this environment more fragile than others. The urban scenario was somewhere in the middle with a mean reliance of 53.81% (SD = 14.93).

For the ANOVA, the type of context acted as the independent variable and the reliance on the automated agent as the dependent one. The results indicate a weak but significant effect of context on the reliance on automation ($F(2, 84) = 4.81$, $p = .011 < .05$, $\eta^2_G = 0.092$). This effect remained significant even after the removal of one outlier that chose not to automate whenever the medical scenario was presented (see Fig. 4). Subsequently performed orthogonal planed contrasts reveal that, in the medical context, the reliance on automation is significantly lower in contrast to the two other conditions ($t(84) = -2.66$, $p = .009 < .05$). The differences in means between the urban and warehouse scenario were not significant. Performing ANOVA with the context as independent variable and the decision times as dependent variable did not lead to any significant differences in means between the different scenarios. Fig. 4 illustrates the mean reliance in dependence on the context.

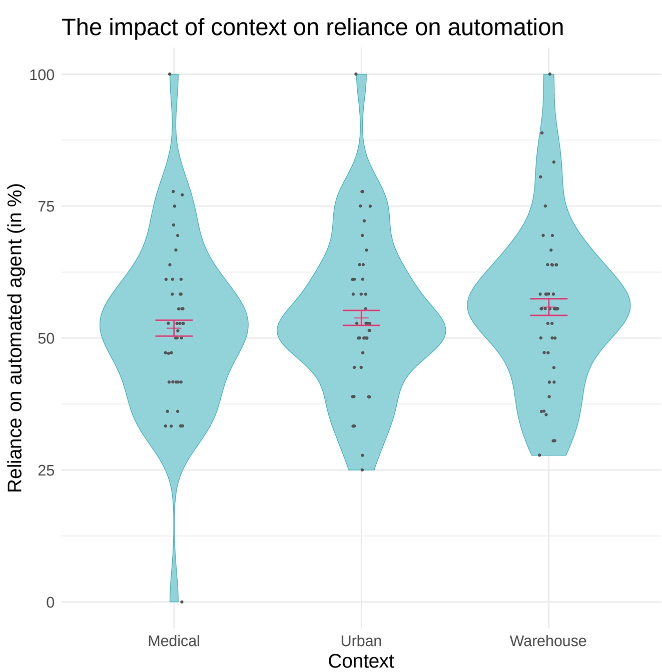


Fig. 4: The impact of context on the reliance of the automated agent including error bars and mean in violin plot.

Again, the statistical results are in line with the analysis of the post-experiment interviews. Participants reported the medical context to be more fragile and that they would therefore be more careful in their decision-making process. They furthermore stated that, in this particular environment, small mistakes would lead to dramatic consequences and that they were less willing to take risks. Additionally, reflecting the absence of differences between the urban and warehouse context, participants noted that they would come to similar decisions in both environments.

Lastly, participants mentioned that the highest amount of risk was taken in the warehouse scenario, as they considered it to be a closed environment in which workers are specifically trained for the purpose of working with automated machines.

Finally, the effect of the context on the perception of safety, the reliance on the previous experience of the automated agent, and the willingness to take risk was analysed. Yet, ANOVA did not show a significant main effect on any of these factors. A marginally significant result was solely found for the impact of context on the reliance on the previous experience of the automated agent ($F(1.35, 55.20) = 2.98$, $p = .078 > .05$, $\eta^2_G = .014$). While in the medical scenario the subjective reliance on the previous experience of the vehicle was 4.43 (SD = 1.04) on average, it was higher in the urban (M = 4.63, SD = 0.85) and warehouse scenario (M = 4.67, SD = 0.82). May that as it be, further investigation is required to examine whether these results hold up in larger samples.

C. *Reliability & Decision Support System (DSS)*

Next, we analysed the effect of additional information that was presented to the participants as further guidance. We first evaluated the effect of the reliability of the automated agent that was presented to the participant as the previous experience of the delivery vehicle (H4). Then, we evaluated the effect of the type of DSS on the compliance with the DSS (H5).

Reliability: For each of the trials, participants were randomly presented with one of three different reliability levels: *60%, 80%, 100%.* Unsurprisingly, the highest reliance on the automated agent was recorded for the control condition 100% reliability with a mean score of 89.42% (SD = 15.71). Both the 60% and 80% condition scored lower than the control condition with 27.11% (SD = 18.99) and 44.91% (SD = 24.43) respectively. The results of ANOVA testing, with the level of reliability as the independent and the reliance on automation as a dependent variable, indicated that the manipulation of the factor *reliability* significantly influenced the reliance on the automated agent ($F(2, 84) = 160.24$, $p < .001$, $\eta^2_G = .88$). Subsequent post-hoc tests revealed a pairwise significant difference in reliance for all three reliability levels ($p < .001$). In terms of decision times, it took participants on

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TTS.2021.3118355, IEEE Transactions on Technology and Society

average 3.36 seconds to reach a decision in the 100% reliability condition, while for the 60% and 80% reliability levels decision speeds were lower. In fact, the average decision time was 5.00 (*SD* = *3.28*) seconds in the 60% condition and 5.14 (*SD* = *2.84*) for the 80% reliability level. Moreover, ANOVA, with the decision time as a dependent variable and the level of reliability as independent variable, showed that the level of reliability had a significant effect on the time it took participants to decide whether to automate or not ($F(2, 84) = 21.86$, $p < .001$, $\eta^2_G = .74$). Here again, subsequent post-hoc testing showed pairwise significant difference in decision times for each reliability level.

DSS: The various DSSs were introduced into the study design as a block factor. Therefore, the subjective compliance with the DSS could be assessed in the post-set questionnaire by asking participants to which extent they followed the recommendation of the DSS presented in the previously completed block. Evaluating these subjective measures did, however, not show any significant difference in means regarding the compliance with the different DSSs. Compliance scores were relatively similar with the AI scoring 49.46% (*SD* =25.24), the engineer reaching a mean of 48.75% (*SD* = 26.29), and the statistical analysis 49.86% (*SD* = 25.33).

D. *Interaction effects*

In a last step, we focused on interaction effects. Results revealed a small but significant interaction between Risk × Reliability ($F(6, 252) = 13.46$, $p < .001$, $\eta^2_G = .05$). Consequently, the participants reacted differently to the different levels of risk depending on the reliability of the automated vehicle. For instance, in the case of possible harm to bystanders, reliance on the automated agent dropped significantly lower in low reliability conditions. In fact, when reliability was 100%, reliance on automation was slightly higher in the personal risk condition (*M* = 83.46, *SD* = 25.59), than in the property and personal risk condition (*M* = 79.97, *SD* = 25.59). In contrast, there was a drop in reliance when reliability levels fell below 100% (see Fig. 5). All other two-way interactions were not significant.

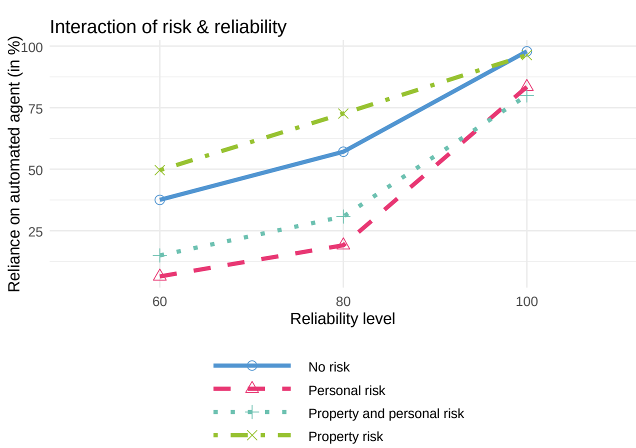


Fig. 5: Interaction effect between the factors risk and reliability on automation reliance.

VI. Discussion

Since automation will continue to increase, it is necessary to address how the willingness to automate is set in different contexts, in situations with moral components or uncertainty, and depending on the availability of information e.g., on reliability. To address these challenges, we designed a game-based study with three different scenarios (warehouse, medical, urban) and with various levels of risk. The controlled gaming environment allowed us to manipulate specific factors known to be as relevant for (moral) decision-making and automation. Our study revealed both expected and unexpected results.

Firstly, our results showed that the decision-making context has a small but measurable impact on the reliance on the automated system, supporting research hypothesis H3. Despite its small effect, this finding was surprising because it is only caused by the visual representation of the task, not the underlying game mechanics. If we assume participant behaviour to be rational, the different contexts should not lead to differences in trust in and reliance on the automated system. We found, nonetheless, that the participants relied less on the automated vehicle in the medical domain, whereas reliance was higher in the warehouse domain. We assume that this was caused by differences in perceived criticality, where accidents were perceived as more severe in the hospital than in the warehouse context (although instructed otherwise). There could also be different basic assumptions about the value of the potential damaged goods or the presence of more persons or objects that might be affected by an accident. In addition, the fact that hospitals are perceived as safe environments for people in a state of vulnerability, suffering injuries in this context might have been considered as extra objectionable. Since we did not actively address these aspects, we cannot rule that they may have been a confound. Additional qualitative studies could provide more insight into the mental models held by our participants, thus allowing for clearer instruction. More knowledge about mental models would also allow us to ask other informative questions about the context of moral decision making, e.g., the impact of risk exposure from the decision-maker's perspective.

Secondly, our study found that participants reacted differently depending on the types of risk associated with automated tasks, supporting research hypothesis H2, namely, that the use of automation was lowest in situations where personal injuries were to be feared and highest when only property damage could occur. No difference was found between the two conditions involving personal injuries alone or in combination with material damages. Our data indicated that the possibility of personal injury alone was sufficient for our participants to refrain from

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TTS.2021.3118355, IEEE Transactions on Technology and Society

using automated systems.

Thirdly, the technical reliability of the automated vehicle had a strong impact on our participants' willingness to, in turn, rely on the vehicle, supporting research hypothesis H4. It is not surprising that reliance declines with lower reliability. However, that even a seemingly perfect automated system is not fully relied upon is an interesting finding as it may indicate that people, in general, may dismiss the benefits of automation even when it is perfectly warranted (*under-reliance* [2], see section 2a). That said, the results of our study are inconclusive in regard to the origin of this effect. Future work should evaluate if and to what extent the following possible explanations may be responsible: Reduced reliance might be shaped by prior negative experiences (c.f., learned distrust from past experimental conditions [54], [70]), a lower overall trust in automated systems in general (based own experience or social learning), or by the presentation of the reliability based on the past performance of the automated agent which may not have been seen as a guarantee for the actual performance.

Fourthly, we could not find any influence of the type of support system on user trust in the support system. Consequently, H5 was rejected. We expected that our participants would attribute different levels of trust to the three different agents, and consequently follow their suggestions to varying degrees. Nevertheless, our results corroborate findings from Clifford and Nass that people may consider interactive media in our three contexts as social agents [71], thus possibly explaining why the expected differences fade. Here, future work with either a significantly larger sample size or more distinguishable representations of the support system (including a digital assistant vs. a human agent) may prove informative.

From a methodological perspective, it is important to note that the research framework we developed using various scenarios was effective as far as the different experimental conditions lead to measurable and consistent changes in attitudes and behaviours in the participants.

Overall, the results and especially the effect of context and risk type on reliance on automation suggest that people trust automation to various degrees depending on where the automation is used and what consequences might arise from failing systems. This implies—at least from the perspective of social acceptance—that automation and machine ethics are not universal and task-independent but that they must be specifically tailored to the task and usage context.

VII. Limitations

While the present study seems to provide robust results, it does not come without its limitations.

First, profits and losses in the game were selected arbitrarily. This is not problematic in this context, as one of the foci lay on the experimental separation of moral and economic decision making. Future studies may vary gains and losses to investigate the relationship between moral considerations and economic theories, e.g. Prospect Theory [72], and examine if users' moral decision-making changes at higher stakes.

Further, we used a scenario-based approach to study how people would interact with automated systems in moral situations. Although people will most likely react differently in real-world environments, prior work from automation research suggests that the findings are partially transferable to real-world decision making [70], especially in regard to the factors that do and do not influence decision making. Nonetheless, future work should study how user preferences change with closer proximity to real decision situations either by providing more contextual information, which could ensure that any uncertainties in assumptions about risk exposure are mitigated, or by interviewing participants in more or even entirely realistic decision-making situations. It also seems advisable to analyse the participants' mental models and decision-making motives related to the investigated scenarios and check to what extent these are an influencing aspect or not.

VIII. Conclusion

The applied research framework described here is one approach to identify and quantify factors crucial to understanding users' expectations of how automation should behave and whether it is implemented within their expectations. This is key as automated systems will only be socially accepted if technical performance is aligned with ethical and legal considerations as well as social expectations [11].

This study suggests that it makes sense to generate more information about the specific characteristics of the stress field of human atomization and morally relevant decisions in the context of controllable environments for decision tasks.

In the near future, it will also be essential to learn more about the factors that influence people's trust in automated systems in the context of moral decisions. More knowledge about the conditions and willingness to have automated systems perform morally sensitive tasks will be crucial for how quickly technologies, such as automated driving, can become established. In addition to the other aspects not examined here, we will have to clarify how to technically implement and communicate machine ethics that, from the human point of view, are also legally justifiable and trustworthy. The design of automated systems will be a success if both technical *and* human requirements are harmonized and well-aligned.

Appendix

A. *Adapted items from Jian et al. (2000)*

TABLE I
Adapted items from the Empirically Determined Scale of Trust in Automated Systems [73]

Original Item	*Adapted item*

I can trust the system.	*In general, I will trust autonomous systems.*
I am wary of the system.	*I am careful when using autonomous systems.*
The system is reliable.	*Autonomous systems are reliable.*
-	*I mistrust decisions of autonomous systems.*
The system's actions will have harmful or injurious consequences.	*The actions of autonomous systems will have negative consequences*

Acknowledgement

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC-2023 Internet of Production—390621612. We thank Monica Gonzalez-Marquez for editing the manuscript.

IX. References

[1] L. Bainbridge, "Ironies of Automation," Automatica, vol. 19, no. 6, pp. 775–779, 1983.

[2] R. Parasuraman and V. Riley, "Humans and Automation: Use, Misuse, Disuse, Abuse," Human Factors: The Journal of the Human Factors and Ergonomics Society, vol. 39, no. 2, pp. 230–253, 1997, doi: 10.1518/001872097778543886.

[3] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans, vol. 30, no. 3, pp. 286–297, 2000, doi: 10.1109/3468.844354.

[4] B. French, A. Duenser, and H. Andrew, "Trust in Automation A Literature Review," CSIRO report EP184082. 2018. Accessed: Nov. 16, 2019. [Online]. Available: http://www.tascl.org/uploads/4/9/3/3/49339445/trust_in_automation_-_a_literature_review_report.pdf

[5] N. Wiener, "Some moral and technical consequences of automation," Science, vol. 131, no. 3410, pp. 1355–1358, 1960, doi: 10.1126/science.131.3410.1355.

[6] N. J. Goodall, "Machine Ethics and Automated Vehicles," Springer, Cham, 2014, pp. 93–102. doi: 10.1007/978-3-319-05990-7_9.

[7] N. J. Goodall, "Ethical Decision Making during Automated Vehicle Crashes," Transportation Research Record: Journal of the Transportation Research Board, vol. 2424, no. 1, pp. 58–65, Jan. 2014, doi: 10.3141/2424-07.

[8] P. Wintersberger, A.-K. Frison, A. Riener, and S. Thakkar, "Do moral robots always fail? Investigating human attitudes towards ethical decisions of automated systems," in 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), Aug. 2017, pp. 1438–1444. doi: 10.1109/ROMAN.2017.8172493.

[9] A.-K. Frison, "Moral Behavior of Automated Vehicles: The Impact on Product Perception," 2018. doi: 10.18420/muc2018-ws15-0475.

[10] P. Foot, "The problem of abortion and the doctrine of double effect.," The Oxford Review, vol. 5, 1967, doi: 10.1002/9781444323528.ch41.

[11] E. Awad et al., "The Moral Machine experiment," Nature, vol. 563, no. 7729, pp. 59–64, 2018, doi: 10.1038/s41586-018-0637-6.

[12] J.-F. Bonnefon, A. Shariff, and I. Rahwan, "The social dilemma of autonomous vehicles," Science, vol. 352, no. 6293, pp. 1573–1576, Jun. 2016, doi: 10.1126/science.aaf2654.

[13] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," International Journal of Human Computer Studies, vol. 58, no. 6, pp. 697–718, 2003, doi: 10.1016/S1071-5819(03)00038-7.

[14] M. Ghazizadeh, J. D. Lee, and L. N. Boyle, "Extending the Technology Acceptance Model to assess automation," Cognition, Technology and Work, vol. 14, no. 1, pp. 39–49, 2012, doi: 10.1007/s10111-011-0194-3.

[15] M. Ghazizadeh, Y. Peng, J. D. Lee, and L. N. Boyle, "Augmenting the Technology Acceptance Model with trust: Commercial drivers' attitudes towards monitoring and feedback," Proceedings of the Human Factors and Ergonomics Society, pp. 2286–2290, 2012, doi: 10.1177/1071181312561481.

[16] B. M. Muir, "Trust in Atomation. Part I. Theoretical Issues in the Study of Trust and Human Intervention in Automated Systems," Ergonomics, vol. 37, no. 11, pp. 1905–1922, 1994, doi: 10.1080/00140139408964957.

[17] B. M. Muir and N. Moray, "Trust in Automation. Part II. Experimental Studies of Trust and Human Intervention in a Process Control Simulation," Ergonomics, vol. 39, no. 3, pp. 429–460, 1996, doi: 10.1080/00140139608964474.

[18] J. D. Lee and K. A. See, "Trust in Automation: Designing for Appropriate Reliance," Human Factors, vol. 46, no. 1, pp. 50–80, 2004, doi: 10.1518/hfes.46.1.50_30392.

[19] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. De Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," Human Factors, vol. 53, no. 5, pp. 517–527, 2011, doi: 10.1177/0018720811417254.

[20] D. R. Billings, K. E. Schaefer, J. Y. C. Chen, and P. A. Hancock, "Human-robot interaction: Developing trust in robots," in HRI'12 - Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction, 2012, pp. 109–110. doi: 10.1145/2157689.2157709.

[21] P. A. Hancock, D. R. Billings, and K. E. Schaefer, "Can you trust your robot?," Ergonomics in Design, vol. 19, no. 3, pp. 24–29, 2011, doi: 10.1177/1064804611415045.

[22] J. K. Rempel, J. G. Holmes, and M. P. Zanna, "Trust in Close Relationships," Journal of Personality and Social Psychology, vol. 49, no. 1, pp. 95–112, 1985, doi: 10.1037/0022-3514.49.1.95.

[23] B. M. Muir, "Trust between humans and machines, and the design of decision aids," International Journal of Man-Machine Studies, vol. 27, no. 5–6, pp. 527–539, 1987, doi: 10.1016/S0020-7373(87)80013-5.

[24] J. D. Lee and N. Moray, "Trust, control strategies and allocation of function in human-machine systems," Ergonomics, vol. 35, no. 10, pp. 1243–1270, 1992, doi: 10.1080/00140139208967392.

[25] J. D. Lee and K. a See, "Trust in Automation: Designing for Appropriate Reliance," Human Factors, vol. 46, no. 1, pp. 50–80, 2004, doi: 10.1518/hfes.46.1.50.30392.

[26] SAE, "SAE International Releases Updated Visual Chart for Its 'Levels of Driving Automation' Standard for Self-Driving Vehicles," 2018.

[27] D. Rushe, "Tesla driver who died in 'autopilot' crash was playing on phone, inquiry finds." 2020. Accessed: Apr. 03, 2021. [Online]. Available: https://www.theguardian.com/technology/2020/feb/25/tesla-driver-autopilot-crash

[28] R. R. Hoffman, M. Johnson, J. M. Bradshaw, and A. Underbrink, "Trust in automation," IEEE Intelligent Systems, vol. 28, no. 1, pp. 84–88, 2013, doi: 10.1109/MIS.2013.24.

[29] M. Lewis, K. Sycara, and P. Walker, "The Role of Trust in Human-Robot Interaction," in Studies in Systems, Decision and Control, vol. 117, Springer International Publishing, 2018, pp. 135–159. doi: 10.1007/978-3-319-64816-3_8.

[30] I. Krumpal, "Determinants of social desirability bias in sensitive surveys: A literature review," Quality and Quantity, vol. 47, pp. 2025–2047, 2013, doi: 10.1007/s11135-011-9640-9.

[31] K. K. Damghani, M. T. Taghavifard, and R. T. Moghaddam, "Decision Making Under Uncertain and Risky Situations," Enterprise Risk Management Symposium Monograph Society of Actuaries - Schaumburg, no. 148, 2009.

[32] T. Vasvári, "Risk, Risk Perception, Risk Management – a Review of the Literature," Public Finance Quarterly, vol. 60, no. 1, pp. 29–48, 2015.

[33] S. Toma and M. Chiti, "Risk and Uncertainty," Procedia Economics and Finance, vol. 3, pp. 975–980, 2012, doi: 10.1016/S2212-5671(12)00260-2.

[34] O. Renn and C. Benighaus, “Perception of technological risk: Insights from research and lessons for risk communication and management,” Journal of Risk Research, vol. 16, no. 3–4, pp. 293–313, 2013, doi: 10.1080/13669877.2012.729522.

[35] M. H. Birnbaum, “Behavioral models of decision making under risk,” in Psychological Perspectives on Risk and Risk Analysis: Theory, Models, and Applications, Springer International Publishing, 2018, pp. 181–200. doi: 10.1007/978-3-319-92478-6_8.

[36] G. Breakwell, The psychology of risk. Cambridge University Press, 2014.

[37] O. Renn, “Mental Health, Stress and Risk Perception: Insights from Psychological Research,” in CIBA Foundation Symposia, no. 203, Novartis Foundation, 2007, pp. 205–234. doi: 10.1002/9780470515006.ch15.

[38] O. Renn, “Risk perception and risk management: A Review. Part 1: Risk perception,” Risk Abstracts, vol. 7, no. 1, pp. 1–9, 1990, doi: 10.18419/opus-7299.

[39] P. Slovic, J. Flynn, C. K. Mertz, M. Poumadère, and C. Mays, “Nuclear Power and the Public: A Comperative Study of Risk Perception in France and the United States,” in Cross-Cultural Risk Perception, Springer US, 2000, pp. 55–102. doi: 10.1007/978-1-4757-4891-8_2.

[40] S. H. Schwartz and W. Bilsky, “Toward a Theory of the Universal Content and Structure of Values: Extensions and Cross-Cultural Replications,” Journal of Personality and Social Psychology, vol. 58, no. 5, pp. 878–891, 1990, doi: 10.1037/0022-3514.58.5.878.

[41] C. W. Scherer and H. Cho, “A social network contagion theory of risk perception,” Risk Analysis, vol. 23, no. 2, pp. 261–267, 2003, doi: 10.1111/1539-6924.00306.

[42] B. Garrigan, A. L. R. Adlam, and P. E. Langdon, “Moral decision-making and moral development: Toward an integrative framework,” Developmental Review, vol. 49, pp. 80–100, 2018, doi: 10.1016/j.dr.2018.06.001.

[43] S. O. Hansson, “Ethical Criteria of Risk Acceptance,” Erkenntnis, vol. 59, pp. 291–309, 2003, doi: 10.1023/A:1026005915919.

[44] M. Hayenhjelm and J. Wolff, “The Moral Problem of Risk Impositions: A Survey of the Literature,” European Journal of Philosophy, vol. 20, no. S1, pp. E26–E51, Jun. 2012, doi: 10.1111/j.1468-0378.2011.00482.x.

[45] N. J. Goodall, “Away from Trolley Problems and Toward Risk Management,” Applied Artificial Intelligence, vol. 30, no. 8, pp. 810–821, 2016, doi: 10.1080/08839514.2016.1229922.

[46] B. Lundgren, “Safety requirements vs. crashing ethically: what matters most for policies on autonomous vehicles,” AI and Society, vol. 36, pp. 405–415, 2020, doi: 10.1007/s00146-020-00964-6.

[47] S. O. Hansson, “Risk,” in Stanford Encyclopedia of Philosophy, Fall 2018., Edward N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2018. Accessed: Nov. 11, 2020. [Online]. Available: https://plato.stanford.edu/entries/risk/#Eth

[48] Y. E. Bigman and K. Gray, “People are averse to machines making moral decisions,” Cognition, vol. 181, pp. 21–34, Dec. 2018, doi: 10.1016/j.cognition.2018.08.003.

[49] J. Gogoll and M. Uhl, “Rage against the machine: Automation in the moral domain,” Journal of Behavioral and Experimental Economics, vol. 74, pp. 97–103, Jun. 2018, doi: 10.1016/j.socec.2018.04.003.

[50] G. A. Gorry and M. Morton, “A framework for management information systems,” MIT Sloan Management Review, 1989.

[51] J. P. Shim, M. Warkentin, J. F. Courtney, D. J. Power, R. Sharda, and C. Carlsson, “Past, present, and future of decision support technology,” Decision Support Systems, vol. 33, no. 2, pp. 111–126, 2002, doi: 10.1016/S0167-9236(01)00139-7.

[52] M. Madsen and S. Gregor, “Measuring Human-Computer Trust,” in Proceedings of Eleventh Australasian Conference on Information Systems, 2000, pp. 6–8.

[53] B. M. Muir and N. Moray, “Trust in Automation. Part II. Experimental Studies of Trust and Human Intervention in a Process Control Simulation,” Ergonomics, vol. 39, no. 3, pp. 429–60, 1996, doi: 10.1080/00140139608964474.

[54] P. de Vries, C. Midden, and D. Bouwhuis, “The effects of errors on system trust, self-confidence, and the allocation of control in route planning,” International Journal of Human Computer Studies, vol. 58, no. 6, pp. 719–735, 2003, doi: 10.1016/S1071-5819(03)00039-9.

[55] B. Brehmer and D. Dörner, “Experiments with computer-simulated microworlds: Escaping both the narrow straits of the laboratory and the deep blue sea of the field study,” Computers in Human Behavior, vol. 9, no. 2–3, pp. 171–184, 1993, doi: 10.1016/0747-5632(93)90005-D.

[56] G. L. Liehner, P. Brauner, A. K. Schaar, and M. Ziefle, “Development and Evaluation of a Research Framework for Measuring the Reliance on Automation in Situations of Risk and Moral Dilemma,” in Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Human Communication, Organization and Work, V. Duffy, Ed. Springer, Cham, 2020, pp. 280–295. doi: 10.1007/978-3-030-49907-5_20.

[57] J. Sanchez, W. A. Rogers, A. D. Fisk, and E. Rovira, “Understanding reliance on automation: Effects of error type, error distribution, age and experience,” Theoretical Issues in Ergonomics Science, vol. 15, no. 2, pp. 134–160, 2014, doi: 10.1080/1463922X.2011.611269.

[58] L. Perkins, J. E. Miller, A. Hashemi, and G. Burns, “Designing for Human-Centered Systems: Situational Risk as a Factor of Trust in Automation,” Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 54, no. 25, pp. 2130–2134, 2012, doi: 10.1177/154193121005402502.

[59] M. Desai et al., “Effects of changing reliability on trust of robot systems,” in HRI’12 - Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction, 2012, pp. 73–80. doi: 10.1145/2157689.2157702.

[60] L. Wang, G. A. Jamieson, and J. G. Hollands, “Trust and reliance on an automated combat identification system,” Human Factors, vol. 51, no. 3, pp. 281–291, 2009, doi: 10.1177/0018720809338842.

[61] E. Rovira, K. McGarry, and R. Parasuraman, “Effects of imperfect automation on decision making in a simulated command and control task,” Human Factors, vol. 49, no. 1, pp. 76–87, 2007, doi: 10.1518/001872007779598082.

[62] A. Chavaillaz, D. Wastell, and J. Sauer, “System reliability, performance and trust in adaptable automation,” Applied Ergonomics, vol. 52, pp. 333–342, Jan. 2016, doi: 10.1016/j.apergo.2015.07.012.

[63] J. Bruner, Industrial Internet: The Machines are Talking. O’Reilly Media, Inc, 2013.

[64] H. Lasi, P. Fettke, H. G. Kemper, T. Feld, and M. Hoffmann, “Industry 4.0,” Business and Information Systems Engineering, 2014, doi: 10.1007/s12599-014-0334-4.

[65] C. Beierlein, A. Kovaleva, C. J. Kemper, and B. Rammstedt, “Eine Single-Item-Skala zur Erfassung von Risikobereitschaft: Die Kurzskala Risikobereitschaft-1 (R-1),” 2014.

[66] C. Beierlein, C. J. Kemper, A. Kovaleva, and B. Rammstedt, “Interpersonales Vertrauen (KUSIV3),” 2014. doi: 10.6102/zis37.

[67] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, “Foundations for an Empirically Determined Scale of Trust in Automated Systems,” International Journal of Cognitive Ergonomics, vol. 4, no. 1, pp. 53–71, Mar. 2000, doi: 10.1207/S15327566IJCE0401_04.

[68] R. A. Fisher, Statistical methods for research workers. Edinburgh: Oliver and Boyd, 1925.

[69] J. Neyman and E. S. Pearson, “The testing of statistical hypotheses in relation to probabilities a priori,” Mathematical Proceedings of the Cambridge Philosophical Society, vol. 29, no. 4, pp. 492–510, 1933, doi: 10.1017/S030500410001152X.

[70] K. A. Hoff and M. Bashir, “Trust in automation: Integrating empirical evidence on factors that influence trust,” Human Factors, vol. 57, no. 3, pp. 407–434, 2015, doi: 10.1177/0018720814547570.

[71] B. Reeves and C. Nass, The media equation: how people treat computers, television, and new media. Cambridge University Press, 1996. doi: 10.1109/MSPEC.1997.576013.

[72] D. Kahneman and A. Tversky, “Prospect theory: An analysis of decision under risk,” in Experiments in Environmental Economics, 2018. doi: 10.2307/1914185.

[73] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, “Foundations for an Empirically Determined Scale of Trust in Automated System,” International Journal of Cognitive Ergonomics, vol. 4, no. 1, pp. 53–71, 2000, doi: 10.1207/S15327566IJCE0401_04.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TTS.2021.3118355, IEEE Transactions on Technology and Society