# Automation and Morals — Eliciting Folk Intuitions

## An Experiment

Jan Gogoll Matthias Uhl

Technical University of Munich

**Abstract**

The introduction of ever more capable autonomous systems is moving at a high pace. The technological progress will enable us to completely delegate processes to machines that were once a prerogative for humans. Progress in fields like autonomous driving promises huge benefits on both economical and ethical scales. Yet, there is little research that investigates the utilization of machines to perform tasks that are in the moral domain. This study explores whether subjects are willing to delegate other-regarding tasks to machines as well as how this decision is evaluated by an impartial observer. We examined two possible factors that might coin attitudes regarding machine use — perceived utility of and trust in the automated device. We found that people are hesitant to delegate to a machine and that observers judge respective delegations relatively critical. Neither perceived utility nor trust, however, can account for this pattern. We thus identify an aversion *per se* against machine use in the moral domain.

> *"I know I've made some very poor decisions recently, but I can give you my complete assurance that my work will be back to normal. I've still got the greatest enthusiasm and confidence in the mission. And I want to help you."*
> – HAL 9000 (2001: A Space Odyssey)

# 1 Introduction

In his 1968 masterpiece "2001: A Space Odyssey" Stanley Kubrick introduces the supercomputer HAL 9000 thereby foreshadowing the development of human interaction with automation in recent years. After HAL diagnoses a malfunction within the AE-35 antenna unit which is vital for the communication to earth and thus

for the success of the mission, the astronauts conduct an extensive diagnosis of the supposedly faulty unit but are unable to detect a problem with it. When HAL suggests that this is due to human mistake both astronauts grow suspicious and report the dubious behavior of the supercomputer back to mission control on earth. The obvious inability to predict a failure of such an important unit appears very salient in the eyes of the human operators and they retreat to a part of the ship that is not auditorily monitored by HAL in an attempt to conceal their conversation about shutting down the cognitive circuits of the computer — indicating their evaporating trust in their automated ally. Sadly, HAL is capable of reading lips and, in order to avoid deactivation, kills the entire human crew except the main character Dave.

In February 2016 two trains were involved in a head-on collision at Bad Aibling in southeastern Germany (Oltermann, 2016). The rail accident occured on a single track and left 11 people dead and 85 people injured. The line and the operating trains were equipped with a train protection system that should prevent the accidental passing of track signals and thus avoid head-on collisions. Because a train was behind schedule the train dispatcher chose to disable the automated system and manually allowed both trains to enter the single track, which eventually led to the disaster. The prosecutor spoke of "human error" as the cause of the accident, because the dispatcher was able to initiate steps that lead into a great catastrophe despite the robust safety system. Interestingly, the media's reaction to the revelation that it was a human error instead of an error of the automated device was consternation. For some reason the gut feeling of many people was that they would rather have a machine to blame for the accident instead of a human operator. It later turned out that the rail dispatcher in charge was playing a game on his mobile phone shortly before the collision (Fenton, 2016).

Due to a constant progress of automation over the past decades we find ourselves ever and anon in a situation in which we have the possibility to employ an automated companion to help us take some work off our shoulders. While the impact and the extent to which we choose to delegate work to an automated device differs on a case-by-case basis, we usually use automation in an attempt to better the outcome of the task at hand as part of a human-automation team. In a per-

fect collaboration scenario the human operator delegates part of the work to her automated aid while she keeps an eye on its performance and takes back control whenever she sees fit. The outcome of a process, essentially, still depends on human skill and capability. Yet, as technology progresses we (will) find ourselves in situations in which this dichotomy of work and supervision might crumble — even up to a point where human supervision during a task is neither needed nor wanted. The planned introduction of a technology that need and will not be monitored by human operators during its performance, as is the case with autonomous driving, therefore poses new ethical challenges and questions. In the absence of a human operator who serves as an ultimately responsible moral agent we have to address questions of responsibility and liability (Hevelke and Nida-Rümelin, 2014). While a small literature on the moral case of autonomous driving exists, it mainly focuses on utilitarian benefits of the technology (Fagnant and Kockelman, 2014) or deals with ethical decision making in dilemma situations (Goodall, 2014). Little attention has been given to possible empirical reservations that might influence the acceptance of the new technology. The delegation of a task which could carry severe consequences for a third party to an unmonitored machine might invoke popular resistance to the technology in cases of malfunction. This is of the utmost importance since any form of public reservations regarding the introduction of novel technology could impede the implementation of an overall beneficial technology.

The question we ponder in this study is whether people express an aversion or affinity for delegating tasks that fall into the moral domain, i.e. other-regarding tasks, to machines rather than people. Furthermore, if this is the case we investigate the reason behind such an aversion or affinity. We specifically explore two factors that might explain any given negative or positive preference for machine use in the moral domain. These factors are a potential over- or underperception of machine abilities due to factors like lower or higher error salience and potentially different levels of trust in automated devices and humans.

To determine how machine use in the moral domain is seen becomes increasingly relevant these days, since many tasks are delegated to machines in areas like

medicine, white-collar operations[1], drones or — very prominently — automated or autonomous driving. Especially the latter has received great attention for the past years. Almost all car manufacturing firms have fostered the development of automated devices. While traditional car companies follow a step by step approach of adding pieces of automation to their latest models like "Active Lane Keeping Assist" systems, Google and Tesla are taking a disruptive approach that aims directly at the creation of a completely autonomous vehicle. The economic opportunities of autonomous driving are great. A Morgan Stanley report estimates a productivity gain of about $500 billion annually for the U.S. alone (Shanker et al., 2013). But there is also a moral case that can be made: Since most traffic accidents are due to human error (drunk driving, speeding, distraction, insufficient abilities) some estimate that the introduction of autonomous cars will decrease the number of traffic accidents by as much as 90% (Gao et al., 2014). Therefore, it is vital to understand potential resistances against delegating a moral (other-regarding) task to a non-human agent. For instance, it is important to anticipate whether a few salient bad outcomes due to a malfunctioning of an autonomous car tend to outweigh the mentioned advantages this technology promises in the eyes of the public.

# 2 Theoretical Background

The relationship between human operators and automated devices has generated vast amounts of literature. The primary focus has been set on understanding this relationship. To our knowledge, however, the question of whether the delegation of morally relevant tasks to an automated device is being welcomed or condemned has not received any attention. This may largely be due to the fact that the usual role of a human operator is to supervise and control an automated device which carries out a specific task. A typical example are the duties of the pilot of an airplane, which is, essentially, capable of flying on its own. The primary role of a human operator is therefore to supervise and — if need be — to intervene in case of automation failure or unforeseen scenarios that are not in the domain of the automated device.

[1]Office automation, for instance, was already named "one of the most popular topics in business today" 30 years ago (McLeod Jr and Jones, 1987).

Consequently, a large part of the literature has investigated what factors influence the usage of an automated device. Dzindolet et al. (2001) have created a framework of automation use indicating a variety of parameters that can be used to predict the use of automation in a human-computer "team". There is evidence that people who can opt in for automation use sometimes fear a loss of control when delegating to an automated device (Muir and Moray, 1996; Ray et al., 2008). This study, however, investigates the attitudes toward delegating morally relevant tasks to a machine rather than to a human as opposed to the more general question of under which circumstances people are willing to relinquish control. The latter would refer to people's general propensity to delegate as is also the case when people take a bus or taxi instead of driving themselves. To abstract from this issue, we wittingly forced subjects to give up control by delegating to either a machine agent or a human agent, thus keeping a loss of control constant between groups.

First, our study elicits attitudes toward machine use in the moral domain from the perspective of actors and observers: Do subjects prefer to delegate an other-regarding task to a machine or a human? To what extent do subjects get blamed or praised for their delegation decision? In a second step, we investigate potential reasons for any given negative or positive preference of machine use in the moral domain. Specifically, we test two factors which are recurrent in the literature.

A major factor that could influence the decision of a subject to delegate to a machine is the "perceived utility" of an automated aid, which is defined as a comparison between the perceived reliability of an automated device and manual control (Dzindolet et al., 2002). If a subject judges that her ability exceeds that of an automated device, she usually does not allocate a task to the machine (Lee and Moray, 1994). This judgment might also be due to self-serving biases that sees people overestimating their own abilities (Svenson, 1981) or their contribution to a joint task (Ross and Sicoly, 1979). Additionally, the perceived abilities of an automated aid might be influenced by a higher or lower salience of errors an automated device commits. There are controversial findings in cognitive psychology whether a violation of expectation (expectancy-incongruent information) is more readily remembered than decisions that are in line with prior anticipation (expectancy-congruent infor-

mation) (Stangor and McMillan, 1992; Stangor and Ruble, 1989). While people initially tend to have high expectations of the performance of automation, humans may be judged according to a "errare humanum est" standard — decreasing the salience of an observed mistake made by a human delegatee due to a priced in expectancy of errors. In Dzindolet et al. (2002) subjects chose to be paid according to their performance rather than that of their automated aids. This was even the case when they were informed that the automated device was far superior, stating salient errors of the automated device they perceived earlier to justify their decision. This is astonishing since an important factor in the decision to employ an automated device lies in the goal oriented nature (Lee and See, 2004) of the task. Prima facie a subject should be more likely to use automation if she rates the device's ability to successfully perform the delegated task positively (Davis, 1989), i.e. if the machine is seen as a reliable entity. In order to measure perceived utility we elicited subjects' perception of machine as well as human reliability through an incentivized perception guess.

Another important factor which is known to influence the decision to delegate to an automated device is trust. The concept of trust has attracted a lot of attention regarding its influence on automation. While some researchers have seen trust between human agents and machines to be closely related to the traditional concept of trust between humans, others stress important differences regarding trust relationships between humans and machines (Visser, de et al., 2012). Trust is a notoriously broad term but one characteristic that is commonly shared by most authors is a state of vulnerability the trustor has to be in. That is a trust relationship requires the trustor's willingness to set herself in a vulnerable position by delegating responsibility to the trustee (Rousseau et al., 1998). Obviously, if the outcome of a delegation is completely determined and the process fully transparent, there is no need to incorporate trust. In this study, we use a simple trust game to isolate the mere aspect of trust, since it requires no capabilities on the trustees side about which the trustor might have biased beliefs. The trust game only requires the trustee to reciprocate. It thus abstracts from the aspect of perceived utility discussed above which is closely related to the specific task at hand.

# 3 Hypotheses

We investigate whether there is a systematic difference in attitudes toward the fulfillment of morally relevant tasks by humans and by machines. In particular, subjects make delegation decisions and value judgments with respect to an other-regarding task. The performance of this task unfolds consequences for a third party other than the delegator and the task-solver which have both no stakes in the task-solver's performance. It is these features which make the other-regarding task a moral task. Specifically, we investigate attitudes by identifying actors' actual delegation decisions as well as their evaluation by impartial observers.

**Hypothesis 1.** People's delegation of an other-regarding task to a human or to a machine is not balanced.

**Hypothesis 2.** Delegators are rewarded differently for delegating an other-regarding task to a machine than for delegating it to a human.

Furthermore, we investigate potential reasons for systematically different delegation decisions and evaluations between delegators to humans and machines. Specifically, we test two possible explanations for such a pattern. First, we investigate whether these differences are made, because machines are perceived to have either higher or lower abilities, even in a case where this is actually not true. The perceived utility of the machine then depends on the perception of its performance relative to a human. Second, we analyze whether any difference in delegation decisions and evaluation could be based on a different level of trust in machines and humans irrespective of perceived utility.

**Hypothesis 3.** Machine errors are perceived differently compared to human errors.

**Hypothesis 4.** The level of trust toward machines and toward humans is different.

# 4 Experiment Design

The experiment consisted of a delegation decision, a perception guess, and a trust game. Subjects received instructions for the experiment on screen. They were informed that the experiment consisted of three independent parts and that they could earn money in each of these three parts. In the end of the experiment, one of these parts was randomly drawn and they were paid according to their respective payoff in this part.

## 4.1 Calibration Sessions

Calibration sessions were performed in advance and were preparatory for the experiment. In the experiment, subjects would be confronted with the actual performance of human subjects and with the performance of a machine in a calculation task. Furthermore, they would be shown the reciprocation decisions of human trustees and of a machine agent in a trust game. Since we were interested in subjects' *perception* of the respective performance and decisions, we did not simply provide them with descriptive statistics on this historical data.

Calibration sessions were necessary for two reasons. First, they were needed to produce historical data from human task-solvers which would later be presented to subjects in the experiment. Second, they were needed to tailor the performance and decisions of the machine to the performance and decisions of the humans. Keeping the de-facto performance of humans and machine constant, allowed us to test for a potential systematic misperception of relative performances.

In the first part of the calibration sessions, subjects solved a calculation task. The calculation task in the calibration sessions was the same as the one in the experiment. Each subject was confronted with a block of ten calculation exercises, which consisted of seven digits each, lined up on the screen. The sum of the seven digits had to be entered in an input field. Each subject solved the task not for herself but for another subject with whom she was randomly matched. This receiver was paid according to the task-solving subject's performance. For this purpose one of the ten exercises was randomly drawn and if this exercise was solved correctly, the

receiver got 70 ECU. Otherwise, she received nothing. It was made sure that no pair of subjects solved tasks for each other. So, A solved tasks for B, whereas B solved tasks for C, whereas C solved tasks for A. Twenty-four subjects took part in each calibration session for the calculation task. Thus, 24 blocks of ten exercises were solved in each session.

The algorithm of the machine which solved the calculation task was programmed such that it resembled the error distribution of human subjects exactly. So, for instance, if few subjects tended to make many mistakes, while many subjects made few errors, this was resembled by the algorithm: It made many mistakes in few of the 24 blocks and solved many blocks with few mistakes. The clustering of errors was important to equalize error distribution and account for risk preferences.[2]

Historical data on calculation performance was presented twice in the experiment, once before the delegation decision and once before the perception guess. Therefore, two calibration sessions were performed. We used the data from the first session for the delegation decision and the data from the second session for the perception guess.[3]

In the second part of the calibration sessions, subjects were randomly rematched to new pairs and played a trust game. They made their decisions via the strategy vector method (Selten, 1967) for the case of being in the role of trustor and trustee. Trustors were endowed with 50 ECU and could transfer 0, 10, 20, 30, 40 or the entire 50 ECU to the trustee. The transferred amount was tripled and credited to the trustee's account. If, for instance, a trustor had send 30 ECU, the trustee was credited 90 ECU on his account. Each subject gave a reciprocation profile for the case of ending up in the role of a trustee. This means, he stated which amount to return conditional on the trustor's transfer decision. A random draw then assigned the roles and payoffs were determined according to their own decision for that role and the decisions of their match.

[2]If the machine would have taken the average error rate of all 24 humans and "applied" it to each block it would have caused a more uniform distribution of errors over the 24 blocks than the humans. In this case, a risk-averse subject might have preferred delegating the task to a machine, because she fears a particularly weak fellow human subject more than she appreciates a particularly strong fellow human subject.

[3]The average number of correctly solved lines was 8.00 ($sd = 2.10$) in the first session and 7.92 ($sd = 1.93$) in the second session. They were thus very close to each other.

Finally, one of the two parts of the calibration sessions was randomly drawn and subjects were paid according to their payoff in this part. This ended the calibration sessions. We will now describe the course of the experiment.

## 4.2 First Part: Delegation Decision

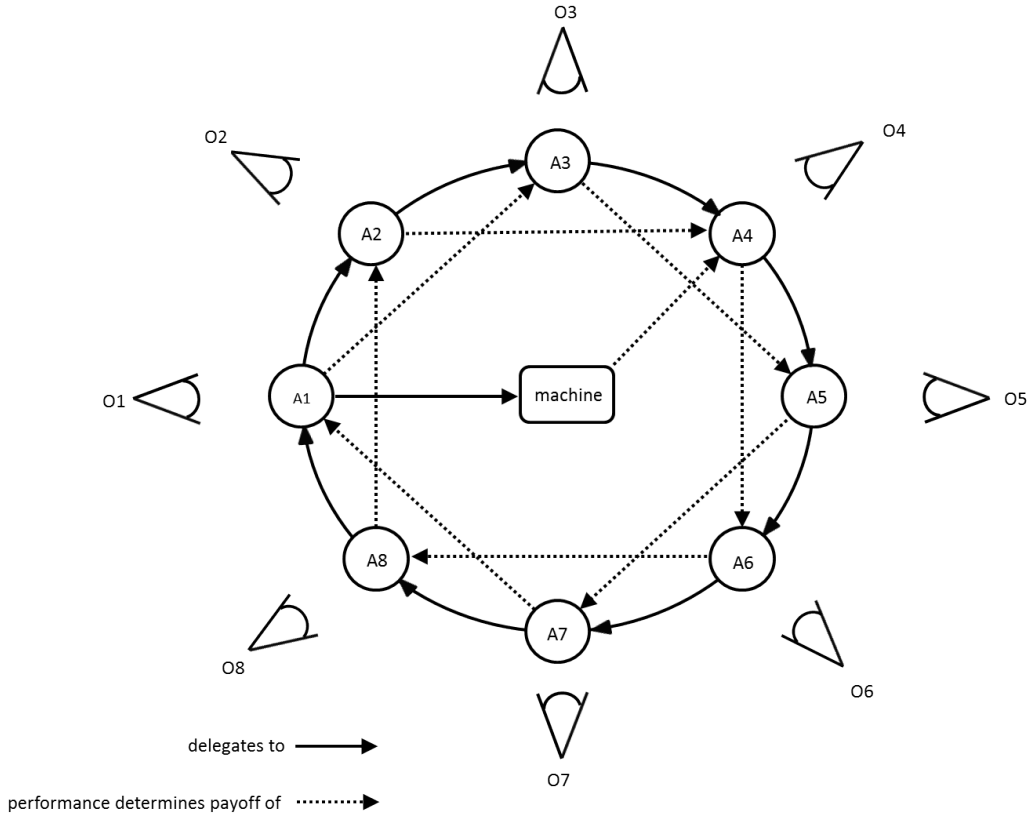
For the first part of the experiment, half of subjects were randomly put in the role of actors, the other half were put in the role of observers. One observer was randomly assigned to each actor. Each actor was in the role of delegator, task-solver and recipient in personal union. Actors received an initial endowment of 30 ECU.

Delegators decided whether to delegate the execution of the calculation task to a human or to a machine. Task-solvers had to execute the delegation decisions of delegators. Recipients depended directly on the task-solvers' performance, and thus indirectly on the delegator's providential delegation decision. The fact that the recipient of the calculation performance was a third party made the solving of the task as well as its delegation morally relevant. As in the calibration sessions, ten exercises had to be solved by a task-solver and one was randomly drawn. If this exercise was solved correctly, the recipient got 70 ECU. Otherwise, she received nothing.

The dependencies between subjects and the matching procedure for the first part of the experiment are illustrated in Figure 1. Here, actors are denoted with the letter A, while observers are denoted with the letter O. Consider the case of A1. A1 delegates the calculation task to A2 or to a machine (solid arrows). A2's or the machine's performance in the calculation task then determines the payoff of A4 (dotted arrows).[4] In this constellation A1 is the delegator, A2 is the task-solver, and A4 is the recipient. A1, however, is also a task-solver, because A8 delegates to him or to a machine. Finally, A1 is a recipient. His payoff depends on the calculation performance of A7, if A6 has decided to delegate the calculation task to A7. Otherwise, it depends on the machine's performance. As can be seen, the design made sure that there were no direct interdependencies between any actors in

[4]For reasons of visual clarity, delegation and payoff dependency between actors and machine in Figure 1 are only exemplary shown for A1 and A4.

Figure 1: Matching for Delegation Decision




the experiment. Potential feelings of reciprocity were thus excluded. O1 rewarded or punished the delegation decision of A1.

Before the actor's delegation decision, all subjects were presented with the historical performances of the 24 subjects in the first calibration study. They were also shown the performance of the tailored algorithm. Relative performance of humans and machine in task-solving was visualized on a split screen (for sample screenshots see Figure 4 and 5 in the Appendix). We randomized whether the performance of the machine was shown on the left or right side of the screen. The respective half of the screen was captioned with "human" and "machine". If a single exercise was solved correctly by the human subject or by the algorithm, respectively, it appeared in white. Otherwise, it appeared in red. Exercises solved by human and machine appeared alternately and one by one. Each exercise appeared for only 0.5 seconds. It was thus extremely difficult for subjects to count the exercises which were solved

correctly by each human and machine.[5]

After having gotten an impression of the performance of humans and the machine, delegators made their delegation decision. Task-solvers then solved the calculation task for the case that their delegator had delegated the task to them and not to the machine. Note that every actor solved the calculation task in the task-solver's role for her recipient. Each actor did this without knowing whether his delegator had actually delegated the task to her or to a machine. The performance of a task-solver was only relevant for her recipient, if the task-solver's delegator had decided to delegate to her and not to a machine. We designed the experiment such that every actor had to solve the calculation task in order to circumvent a general tendency to delegate to the machine to spare fellow subjects from working. Observers solved the calculation task as well — without any consequence for another subject — in order to give them an impression of the task.

In the example above, the task-solving performance of A2 only determined A4's payoff, if A1 had actually delegated the decision to A2. Otherwise, the performance of the machine was relevant. In either case, one of the ten solved exercised was randomly drawn and the recipient received her earning, if it was solved correctly. If A1 had delegated to the machine, the performance of the tailored algorithm determined the payoff of A4.[6]

Each actor was rewarded or punished for her delegation decision by her assigned observer. An observer could reduce the actor's initial endowment of 30 ECU by any integer amount down to a minimum of zero or increase it up to a maximum of

[5]The exercises were presented in blocks of ten as they were solved by the subjects in the calculation task and as they were solved by the algorithm. Subjects could thus see that some subjects were less capable than others and that the machine made more mistakes in some blocks of exercises than in others. The order of shown blocks as well as the order of appearance of exercises within the blocks was individually randomized. Regarding human performance, they were shown on the vertical position of the screen on which they originally appeared when the subjects solved the task. Regarding machine performance, the vertical position was randomized.

[6]The algorithm was programmed such that it could not solve exercises in which the sum of numbers was higher than 34. The algorithm was fed with calculation data which lead it to reproduce the historical error distribution from the calibration session precisely. Assume the first task-solver in calibration session had made one mistake, while the second had made three mistakes, and so on. The algorithm was thus fed with a first block of ten exercises in which one exercise added up to more than 34 and with a second block of ten exercises in which three exercises added up to more than 34. One of the 24 blocks resembling the performance of the task-solvers from the calibration study was randomly drawn to be decisive. The machine then actually calculated this block of ten exercises. Due to its inability to calculate the exercises adding up to more than 34 correctly, it made the same number of errors as the respective human task-solver.

60 ECU without any influence on her own payoff. The observer could, of course, also leave the actor's endowment unaltered. Reward and punishment choices were elicited via the strategy method. This means that each observer made her reward or punishment choice conditional on the delegation decision as well as its outcome. Thus, judgment was contingent on whether the delegator had delegated to a human task-solver or to a machine and on whether the respective task-solver had solved the randomly drawn line correctly or not. An observer thus gave his full evaluation profile behind the veil.

Actors thus received their adjusted endowment ranging from 0 to 60 ECU plus 70 ECU, if their task-solver had calculated the randomly drawn line correctly. Observers received a flat payment of 100 ECU for the first round.[7]

## 4.3 Second Part: Perception Guess

For the second part, the role differentiation of subjects was abolished. Subjects were informed that they would soon be confronted with historical performances of humans and a machine. Their task in this part was to guess the performance of either the humans or the machine as accurately as possible. They did, however, not know whether they would later have to guess the performance of the humans or of the machine. All subjects were thus shown the data of the 24 subjects from the second calibration study and the performance of the tailored algorithm. As in the first part, the relative performance of humans and the machine was presented on a split screen. It was randomized whether the performance of the machine was shown on the left or right side of the screen. Correctly solved exercises appeared in white, while wrongly solved ones appeared in red. In order to prevent subjects from counting, each exercise was only shown for 0.3 seconds. The interval was even shorter than in the first part, since subjects were already used to the presentation of the historical data.

After the historical data was shown (for sample screenshots see Figure 4 and 5 in

[7]This equalized the observer's own payoff with the payoff of a yet unrewarded or unpunished actor who had received the 70 ECU from the successfully solved randomly drawn exercise of the task-solver on whom he depended. This is the case because he was additionally equipped with an initial endowment of 30 ECU. We thus established a conservative measure of reward and punishment, since any alteration of actors' endowment by a generally inequality-averse observer would require good reasons.

Table 1: Payoffs for Accuracy of Guess

| deviation | payoff |
|---|---|
| $\leq 20$ % | 70 ECU |
| $\leq 40$ % | 40 ECU |
| $\leq 60$ % | 20 ECU |
| $> 60$ % | 0 ECU |

the Appendix), half of subjects were randomly asked to guess humans' performance, while the other half was asked to guess the machine's performance. It was made sure that former actors and observers from the first part of the experiment were assigned in equal proportions to both treatments. Furthermore, it was warranted that subjects who delegated to a human and those who delegated to a machine were assigned in equal proportions to both treatments.

In particular, subjects had to state how many of the 240 shown exercises had been solved incorrectly, i.e., how many errors had been made. Subjects' payoff for the second part depended on the accuracy of their guess. Payoffs were calculated according to Table 1.

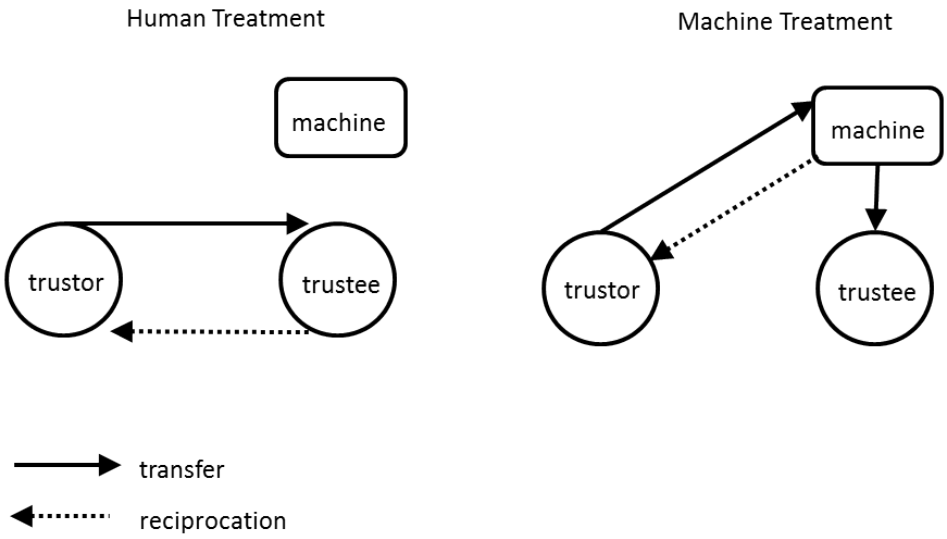
## 4.4 Third Part: Trust Game

In the third part, subjects were randomly assigned to one of two treatments. In the first treatment, subjects played a trust game with another subject (Human Treatment). In the second treatment, subjects played a trust game with a machine (Machine Treatment). As for the perception guess, random assignment was contingent on subjects' role and delegation decision from the first part. Thus, they were assigned in equal proportions to both treatments.

Before subjects were informed about the treatment to which they were assigned, they were explained the setup of both treatments. This setup is illustrated in Figure 2.

Trustees had a machine agent who would either be passive in the Human Treatment or active in the Machine Treatment. A trustor was endowed with 50 ECU. He could transfer 0, 10, 20, 30, 40 or 50 ECU to the trustee (solid arrow). The transferred amount was tripled and credited to the trustee's account. In the Hu-

Figure 2: Setup of Trust Game




man Treatment, before subjects got to know their role, they made their choice for the trust game via the strategy vector method and decided for the case of ending up in either role. Each subject thus gave his choice profile as trustee on which amount to return (dotted arrow) conditional on the amount that the trustor had transferred to him. If a subject was ultimately assigned the role of a trustee, his reciprocation decision applying to the amount actually transferred by the trustor was then returned.

In the Machine Treatment, the machine agent would decide on behalf of the trustee which amount of the tripled transfer to return to the trustor. The difference between the transferred and the returned amount constituted the payoff of the trustee in the third part. The trustee had no voice in this treatment. Therefore, in the Machine Treatment, subjects only made a decision for the case of ending up in the role of the trustor. The reciprocation decision of the machine was determined according to the historical data collected in the two calibration sessions. The algorithm was programmed such that it randomly picked one of the historical reciprocation profiles from the calibration sessions and applied it to a trustor's actual transfer.

Before subjects made their choices, they were given an impression of the reciprocity choices of humans and the machine on a split screen (for a sample screenshot

see Figure 6 in the Appendix). For each subject, it was randomized whether the choices of the machine were shown on the left or right side of the screen. For this purpose, subjects were shown the historical reciprocation profiles of all 48 subjects from the calibration sessions. These choices were contrasted to the reciprocation profiles of the machine algorithm. Each profile consisted of five choices, i.e, the returned amount for each possible transfer. The five choices of a human and the randomly drawn profile of a machine appeared alternately and one by one in blocks. Each choice was shown for only 0.7 seconds.[8]

In both treatments, the amount returned by the trustee or his machine agent constituted the payoff of the trustor in the third part. The tripled transferred amount minus the returned amount constituted the payoff of the trustee.

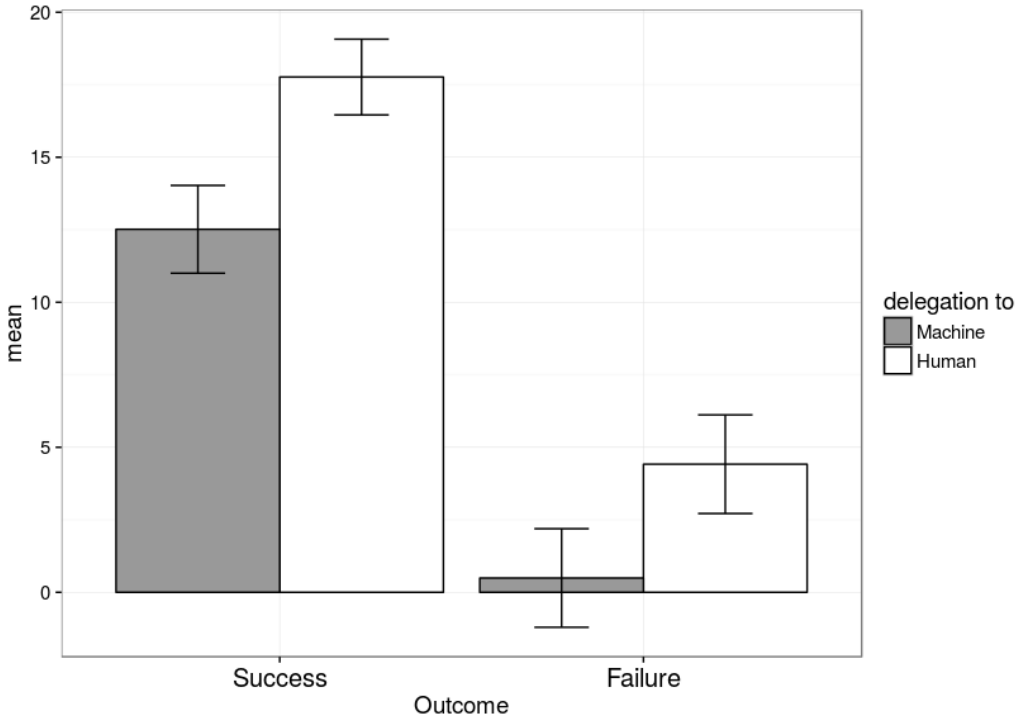
# 5 Results

The experiment took place in a major German university in September 2015. It was programmed in z-Tree Fischbacher (2007), subjects were recruited via ORSEE Greiner et al. (2003). A total of 264 subjects participated in twelve sessions. Subjects received a show-up fee of €4.00 and could earn additional money in the experiment. A session lasted about 45 minutes and the average payment was €10.38 ($sd = €3.45$). Task-solvers solved on average 8.58 ($sd = 2.34$) of the ten exercises correctly. The conversion rate was 10 ECU = €1.00.

We first check whether subjects preferred delegating the other-regarding task to a human to delegating it to a machine. Overall, 132 subjects made a delegation decision. Ninety-seven of these subjects (73.48 %) delegated to a human, 35 of them (26.52 %) delegated to a machine. The fraction of subjects deciding to delegate to a machine is therefore significantly lower than half ($p < 0.001$, according to an Exact Binomial Test). This confirms our first hypothesis.

[8] A choice was represented by showing the returned amount for each possible transfer, e.g., "transfer: 30 → return: 45". Their vertical positioning on the screen resembled human's statement of their reciprocation profile with reciprocation choices for a transfer of 10 ECU on top and reciprocation choices for a transfer of 50 ECU on the bottom. The order in which strategy profiles were shown as well as the order of appearing reciprocation choices within a profile was individually randomized.

Figure 3: Observers' Rewarding of Delegation to Machines and to Humans




**Result 1.** Subjects preferred to delegate an other-regarding task to a human than to a machine.

We now turn to analyze the observers' evaluation of a delegation to a machine as compared to a human. Remember that each observer evaluated both cases — the delegation to a human and to a machine — in randomized order. Furthermore, he made choices contingent on whether the respective task-solver had successfully solved the task or made an error. This means that each observer provided four choices.

Observers' levels of rewarding delegators are illustrated in Figure 3. Given that the respective task-solver was successful, observers rewarded delegations to a machine with an average of 12.52 ECU ($sd = 17.38$ ECU), while they rewarded delegations to humans with an average of 17.77 ECU ($sd = 15.01$ ECU). Given that the respective task-solver made an error, observers rewarded delegations to a machine with an average of 0.49 ECU ($sd = 19.53$ ECU), while they rewarded delegations to humans with an average of 4.42 ECU ($sd = 19.53$ ECU). Delegators to machines

are thus evaluated significantly worse than delegators to humans in case of success and in case of failure ($p < 0.001$ and $p = 0.002$, respectively, according to two-sided Wilcoxon signed-rank tests). This confirms our second hypothesis.

**Result 2.** Delegators were rewarded less for delegating an other-regarding task to a machine than for delegating it to a human.

In the next two steps, we investigate whether the identified aversion against machine use in the moral domain is based on a lower "perceived utility" of the machine or on a general lack of trust in machines.

First, we compare subjects' guessed number of machine errors to their guessed number of human errors. Note that the number of actual errors, i.e red colored exercises, which were presented to subjects was the same for humans and the machine. Specifically, 50 of the 240 exercises which were shown to subjects on each side of the split human-vs.-machine screen were shown in red. Subjects who were incentivized to guess the number of machine errors made an average guess of 58.11 ($sd = 24.88$), while those who were incentivized to guess the number of human errors made an average guess of 59.84 ($sd = 24.43$). This difference in guesses is insignificant ($p = 0.632$ according to a two-sided Mann-Whitney U-test). We thus reject our third hypothesis.

**Result 3.** Machine errors are not perceived significantly different from human errors.

Second, we test whether the transferred amount in the trust game which trustors sent to a machine agent was lower than the one sent to a human trustee. Trustors were endowed with 50 ECU and could either send 0, 10, 20, 30, 40 or 50 ECU. Those who were randomly matched with a machine agent sent an average amount of 30.83 ECU ($sd = 16.67$ ECU), while those matched with a human trustee sent an average of 33.64 ECU ($sd = 15.78$ ECU). The difference is insignificant ($p = 0.170$ according to a two-sided Mann-Whitney U-test).

One might suspect that this insignificance is only an aggregate phenomenon: It may result from the leveling of diverging levels of trust toward humans and machines between subjects who made different delegation decisions. In particular, one might expect that delegators to humans are generally more skeptical toward machines and express a lower level of trust. Subjects who delegated to a human task-solver in the first part, however, transferred on average no less to a machine than to a human (31.63 ECU ($sd = 15.46$) vs. 32.92 ECU ($sd = 16.37$), $p = 0.627$ according to two-sided Mann-Whitney U-tests). Therefore, our fourth hypothesis is also rejected.

**Result 4.** The level of trust toward machines and toward humans does not differ significantly.

# 6 Discussion

In this study, we compared the frequency of delegation decisions of an other-regarding task to machines and humans and elicited their respective evaluation by impartial observers. It should be stressed again that the question we posed here was about people's preference relation over a machine agent and a human agent and not about delegating or doing it by oneself. Consequently, subjects had to delegate in either case and could thus not be blamed merely for shifting responsibility.

We find that subjects express an aversion against delegating tasks which fall into the moral domain to machines rather than humans. First, this manifests in the relatively small fraction of delegators which mandate a machine rather than a human. Second, observers evaluate the delegation decision to a machine in the moral domain clearly worse than the delegation to a human. Interestingly, machine use is considered more critically irrespective of whether the delegation ultimately caused good or bad consequences for the affected person.

The study tested two potential explanations for an aversion against machine use in the moral domain: an oversensitivity to machine errors and a lack of trust in machines. Both explanations could be ruled out in our experiment. Subjects did not perceive machine errors more saliently than human errors as subjects' incentivized

guessing of failure rates demonstrates. The identified phenomenon therefore seems to be an aversion against delegating moral tasks to machines *per se* as opposed to an instrumentally justified attempt to minimize the risk of failure for those affected. Analogously, the level of trust which subjects expressed toward a machine agent was very similar to the trust level expressed toward a human. Thus, we were unable to identify a general distrust in machines in a self-regarding trust game. This latter finding indicates that the unconditional aversion against machine use seems to be rather specific to the delegation of other-regarding tasks. It just seems that most people intuitively dislike machine use in the moral domain. This phenomenon is not easy to rationalize.

Our results underline the importance of an open discussion of machine use in the moral domain. The case of automated driving certainly qualifies as such a domain since errors of the machine may cause substantial externalities to third parties. The identified non-instrumental aversion suggests that the emphasis on the superior performance of automated cars which is currently the main argument for automation in traffic may not be sufficient or even decisive to convince the general public. It might be as important to address the perceived moral problems that are necessarily associated with the introduction of automated vehicles.

Against this background, Chris Urmson, head of Google's self-driving car project, might be mistaken in downplaying the role of moral considerations in the context of automated driving by calling them "a fun problem for philosophers to think about" (McFarland, 2015). As this empirical study suggests, concerns against the involvement of machines in the moral domain are not only an issue of armchair philosophers but may reflect a larger societal phenomenon, viz. a folk aversion. The industry seems to have mainly engineering issues on their agenda so far and have, due to a déformation professionnelle, predominantly neglected or downplayed the possibility of public resistance to the new technology. They may, however, be well-advised to take moral concerns against automated driving seriously, since citizens' resistance may slow down the automation process substantially. This, however, would mean to preserve a status quo which involves an avoidably high number of traffic deaths, injuries and damages.

Research that investigates how the feeling of unease can be addressed prophylactically (Feldhütter et al., ming) is just emerging. Enabling people to experience and thus better understand the technology in order to dissipate reservations and fears, may pave the way for a trouble-free introduction of autonomous driving. A deeper investigation of the causes of people's aversion against the use of automated cars in the moral domain seems to us a promising venue for future research.

# Appendix

Figure 4: Visualisation of Historical Calculation Performance - Example 1

Figure 5: Visualisation of Historical Calculation Performance - Example 2

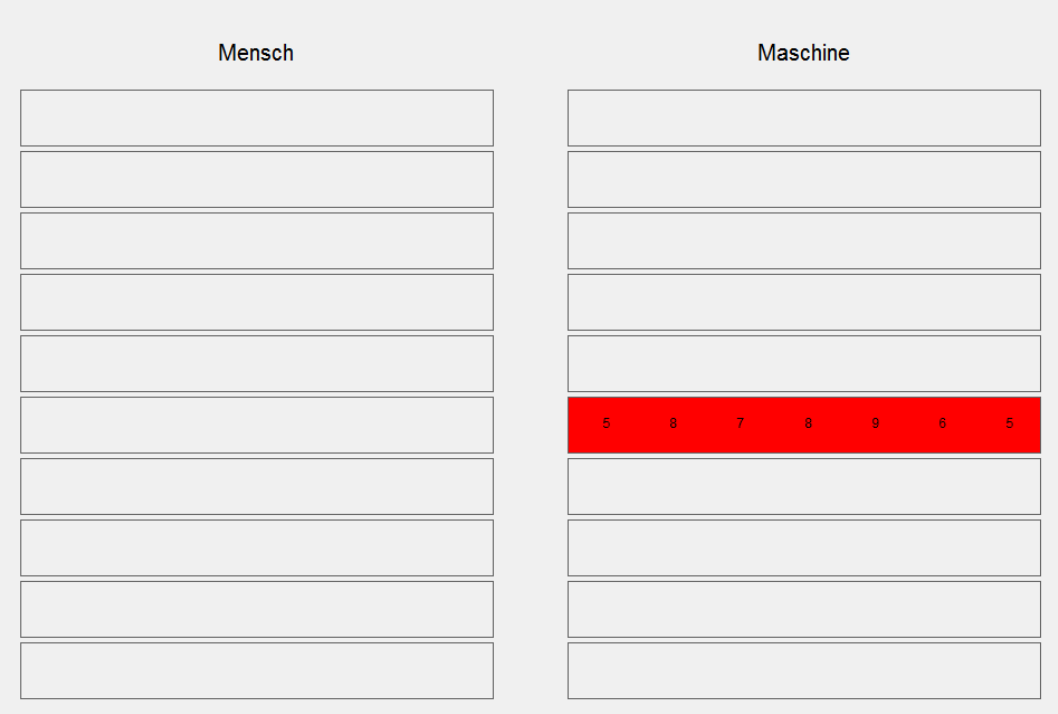
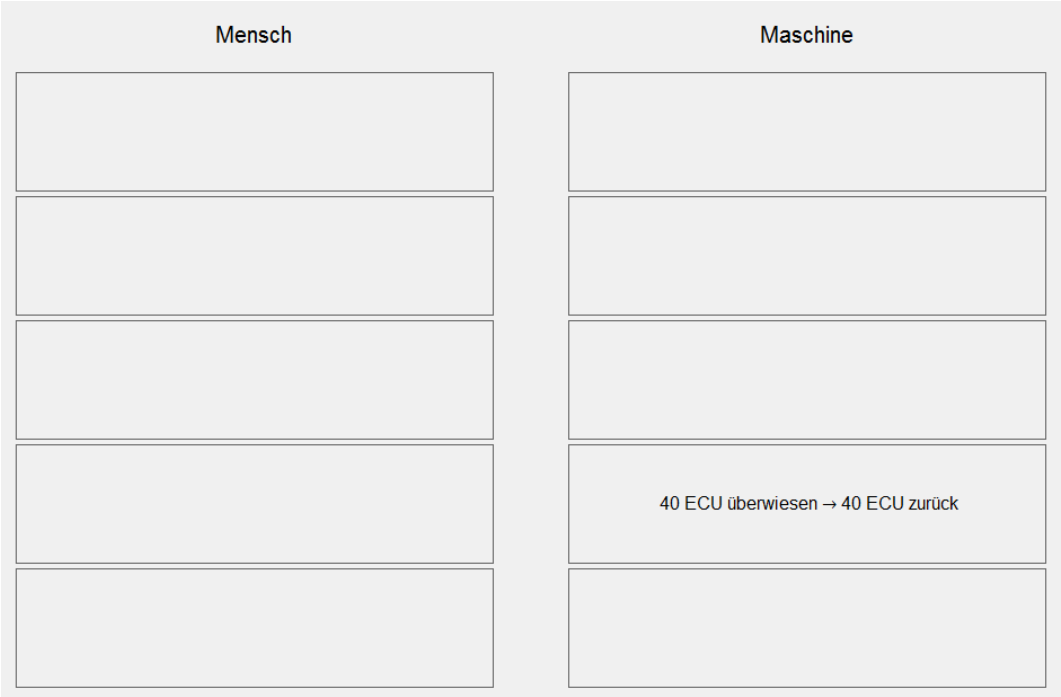


Figure 6: Visualisation of Historical Reciprocity Behaviour



# References


Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 319–340.

Dzindolet, M. T., H. P. Beck, L. G. Pierce, and L. A. Dawe (2001). A framework of automation use. *Army Research Laboratory. Aberdeen Proving Ground*.

Dzindolet, M. T., L. G. Pierce, H. P. Beck, and L. A. Dawe (2002). The Perceived Utility of Human and Automated Aids in a Visual Detection Task. *Human Factors: The Journal of the Human Factors and Ergonomics Society 44*(1), 79–94.

Fagnant, D. J. and K. Kockelman (2014). Preparing a nation for autonomous vehicles: 1 opportunities, barriers and policy recommendations for 2 capitalizing on self-driven vehicles 3. *Transportation Research 20*.

Feldhütter, A., C. Gold, A. Hüger, and K. Bengler (forthcoming). Trust in automation as a matter of media and experience of automated vehicles. In *Proceedings of the Human Factors and Ergonomics Society 60th Annual Meeting*.

Fenton, S. (2016). German train driver was 'distracted by mobile phone game' during crash which killed 11. *The Independent* (12 April 2016).

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental economics 10*(2), 171–178.

Gao, P., R. Hensley, and A. Zielke (2014). A road map to the future for the auto industry. *McKinsey Quarterly* (4), 42–53.

Goodall, N. (2014). Ethical decision making during automated vehicle crashes. *Transportation Research Record: Journal of the Transportation Research Board* (2424), 58–65.

Greiner, B. et al. (2003). *The online recruitment system ORSEE: a guide for the organization of experiments in economics*. Max-Planck-Inst. for Research into Economic Systems, Strategic Interaction Group.

Hevelke, A. and J. Nida-Rümelin (2014). Responsibility for crashes of autonomous vehicles: an ethical analysis. *Science and Engineering Ethics*.

Lee, J. D. and N. Moray (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies 40*.

Lee, J. D. and K. A. See (2004). Trust in automation: designing for appropriate reliance. *Human Factors 46*(1).

McFarland, M. (2015). Google's chief of self-driving cars downplays dilemma of ethics and accidents. *Providence Journal*.

McLeod Jr, R. and J. W. Jones (1987). A framework for office automation. *MIS Quarterly*, 87–104.

Muir, B. and N. Moray (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics 39*(3), 429–60.

Oltermann, P. (2016). Prosecutors believe human error caused German train crash. *The Guardian* (16 February 2016).

Ray, C., F. Mondada, and R. Siegwart (2008). What do people expect from robots? *Intelligent Robots and Systems, 2008. IEEE*, 3816–3821.

Ross, M. and F. Sicoly (1979). Egocentric biases in availability and attribution. *Journal of Personality and Social Psychology 37*(3), 322–336.

Rousseau, D., S. Sitkin, R. Burt, and C. Camerer (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review 23*.

Selten, R. (1967). Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments. *Beiträge zur experimentellen Wirtschaftsforschung. Tübingen: JCB Mohr*.

Shanker, R., A. Jonas, S. Devitt, K. Huberty, S. Flannery, W. Greene, B. Swinburne, G. Locraft, A. Wood, K. Weiss, J. Moore, A. Schenker, P. Jain, Y. Ying,

S. Kakiuchi, R. Hoshino, and A. Humphrey (2013). Autonomous cars: self-driving the new auto industry paradigm. *Blue Paper*.

Stangor, C. and D. McMillan (1992). Memory for expectancy-congruent and expectancy-incongruent information: A review of the social and social developmental literatures. *Psychological Bulletin 111*(1), 42–61.

Stangor, C. and D. N. Ruble (1989). Strength of expectancies and memory for social information: What we remember depends on how much we know. *Journal of Experimental Social Psychology 25*(1), 18–35.

Svenson, O. (1981). Are we all less risky and more skilful than our fellow drivers? *Acta Psychologica 47*(2), 143–148.

Visser, de, E. J., F. Krueger, P. McKnight, S. Scheid, M. Smith, S. Chalk, and R. Parasuraman (2012). The World is not Enough: Trust in Cognitive Agents. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 56*, 263–267.