Sharing Responsibility with a Machine*

Oliver Kirchkamp,[†] Christina Strobel[‡]

January 9, 2019

Humans make decisions jointly with others. They share responsibility for the outcome with their interaction partners. Today, more and more often the partner in a decision is not another human but, instead, a machine. Here we ask whether the type of the partner, machine or human, affects our responsibility, our perception of the choice and the choice itself. As a workhorse we use a modified dictator game with two joint decision makers: either two humans or one human and one machine. We find no treatment effect on perceived responsibility or guilt. We also find only a small and insignificant effect on actual choices.

Keywords: Human-computer interaction; Experiment; Shared responsibility; Moral wiggle room.

JEL Classifications: C91, D63, D80

*This document has been generated on January 9, 2019, with R version 3.5.2 (2018-12-20), on amd64-portbld-freebsd12.0. We thank the Max Planck Society for financial support through the International Max Planck Research School on Adapting Behavior in a Fundamentally Uncertain World (IMPRS Uncertainty). We also thank the audience of IMPRS Uncertainty doctoral seminars and an anonymous referee for useful comments and suggestions. We use R (2018) for the statistical analysis. Data and methods are available at `https://www.kirchkamp.de/research/shareMachine.html` or `http://christina-strobel.de/share_responsibility.html`.

†FSU Jena, School of Economics, Carl-Zeiss-Str. 3, 07737 Jena, `oliver@kirchkamp.de`.

‡FSU Jena, School of Economics, Bachstraße 18k, 07737 Jena, `Christina.Strobel@uni-jena.de`.

1. Introduction

In more and more areas of life decisions are the result of interactions between humans and machines. We encounter automated systems no longer only in a supportive capacity but, more frequently, as systems taking actions on their own. Computer-assisted cars drive autonomously on roads.[1] Surgical systems conduct surgeries independently.[2] We stand no longer on the precipice of a technological change; we are right in the middle of that change. To an average person this becomes easily visible in the area of self-driving cars. The large research investments by car manufacturers as well as tech companies in the last years are only one clear signal for the rapid development of automated system technologies. In 2015 Toyota built a new research institute in Silicon Valley for $1 billion.[3] In 2017 Intel performed its second largest acquisition in the companies history by spending $15.3 billion to buy a company producing camera systems to detect speed limits and potential collisions.[4] In the same year, Ford put $1 billion on the table to acquire Argo AI, an artificial intelligence start up.[5] In 2018, Honda invested $2.75 billion to take a stake in GM Cruise, General Motor's self-driving company.[6] In addition, the jurisdiction is taking an interest in the development of autonomous vehicles. The number of states in the U.S. giving the nod to test self-driving cars on the roads in constantly increasing.[7] In Germany, the government has passed legislation to allow the deployment of automated driving systems in traffic in 2017.[8] Even if the technology has not reached market maturity yet, the top 11 global car manufacturers expect self-driving cars on the highway by 2020 and in urban areas by 2030.[9] A recent report by Navigant Research, a market research institute, forecasts that from 2020 to 2035 around 129 million autonomous vehicles will be sold.[10] Another study on the autonomous vehicle market by Allied Market Research concludes that in 2019 the global market for autonomous vehicles will be worth $54.23 billion and increase by a factor of 10 up to $556.67 billion by 2026.[11]

The next step towards fully autonomous-capable vehicles will be the launch of partly automated driving systems, i.e. cars where the driver hands over the steering or acceleration functions to the vehicle occasionally but can still take control over the vehicle. In all these

[1] See for example, the Tesla car with full self-driving hardware or the NVIDIA AI car that learns from human behavior by using an machine learning approach.

[2] Shademan et al. (2016) also reports a soft tissue surgery conducted by an autonomous system.

[3] See `https://newsroom.toyota.co.jp/en/detail/10171645`.

[4] See `https://newsroom.intel.com/news-releases/intel-mobileye-acquisition/`.

[5] See `https://media.ford.com/content/fordmedia/fna/us/en/news/2017/02/10/ford-invests-in-argo-ai-new-artificial-intelligence-company.html`.

[6] See `https://global.honda/newsroom/news/2018/c181003eng.html`.

[7] See `https://www.brookings.edu/blog/techtank/2018/05/01/the-state-of-self-driving-car-laws-across-the-u-s/`.

[8] See `https://www.freshfields.com/en-gb/our-thinking/campaigns/digital/internet-of-things/connected-cars/automated-driving-law-passed-in-germany/`.

[9] See `https://www.techemergence.com/self-driving-car-timeline-themselves-top-11-automakers/`.

[10] See `https://www.navigantresearch.com/news-and-views/129-million-autonomous-capable-vehicles-are-expected-to-be-sold-from-2020-to-2035`.

[11] See `https://www.alliedmarketresearch.com/autonomous-vehicle-market`.

environments humans find themselves confronted with a new situation: they share decisions with a machine. We call such a situation a hybrid decision situation.[12]

In this paper, we investigate human decision-making in a hybrid decision situation. More specifically, we investigate whether sharing a decision with a computer instead of with another human influences the perception of the situation, thus affecting human decisions. Human decision-making in groups with other humans has been researched extensively. Fischer et al. (2011) show in their meta-study on the so-called bystander effect that the perceived personal responsibility is lower when others are around.[13] Theoretical work from Battigalli and Dufwenberg (2007) and Rothenhäusler et al. (2013) also suggests that people feel less guilty for an outcome when a decision is shared. Furthermore, a meta-study by Engel (2011), including 255 experimental papers on behavior in Dictator Games[14] shows that people behave more selfishly if a decision is shared. So far, however, the literature has only focused on decisions shared between humans. Here we ask whether humans also perceive themselves to be less responsible and guilty and behave more selfishly when the decision is shared with a computer.

As a workhorse, we use a binary Dictator Game. We compare three treatments: a Dictator Game with a single human dictator, a Dictator Game with two human dictators, and a Dictator Game with one human dictator and a computer.

The remainder of the paper is organized as follows: Section 2 provides a literature review focusing on experimental evidence from economics and social psychological research. We especially discuss the literature on individual behavior in groups as well as findings from research on human-computer interactions. In Section 3 we present our experimental design and explain our treatments in more detail. Section 4 relates the experiment to the theoretical background and derives behavioral predictions. Results are presented in Section 5. The last section offers a discussion and some concluding remarks.

2. Review of the literature

In Section 2.1 below, we present previous research on individual decision-making in groups most similar to our experiment. We point out studies explaining why humans behave more selfishly when deciding with other humans. In Section 2.2, we turn to research on human-computer interactions. We outline what is already known about how machines are perceived

[12]However, machines do not always perform better than humans and are also susceptible to errors. The 2016 Disengagement Reports, reports of autonomous vehicle incidents on California public road that all manufacturers in California have to provide to the State of California Department of Motor Vehicles, state 2665 cases in which the test driver had to disengage the autonomous mode (see `https://www.dmv.ca.gov/portal/dmv/detail/vr/autonomous/disengagement_report_2016`) In an international survey about an automized urological surgery by Kaushik et al. (2010) 56.8% of 176 responding surgeons reported to have experienced an irrecoverable intraoperative malfunction of the robotic system.

[13]The *bystander effect*, first described by Latané and Nida (1981), is a social psychological phenomenon that individuals are less likely to help a victim if others are present.

[14]The standard Dictator Game consists of two individuals. One individual – known as the *dictator* – is given some money. The dictator then has to decide how much of this money he/she wants to share with the other individual. The other individual – called the *recipient* – has to accept any amount of money the dictator proposes.

and how humans behave towards them.

2.1. Shared decision-making with humans

People frequently have to make decisions in situations wherein the outcome not only depends on their choice but also the choices of others. In a number of experimental games, such as the Trust Game (Kugler et al., 2007), the Ultimatum Game (Bornstein and Yaniv, 1998), the Coordination Game (Bland and Nikiforakis, 2015), the Signaling Game (Cooper and Kagel, 2005), the Prisoners Dilemma (McGlynn et al., 2009), the Gift Exchange Game (Kocher and Sutter, 2007), the Public Good Games (Andreoni and Petrie, 2004) as well as in lotteries (Rockenbach et al., 2007) and Beauty Contests (Kocher and Sutter, 2005; Sutter, 2005), people have been found to behave more selfish, less trustworthy and less altruistic towards an outsider when deciding together with others.

Even in a game as simple as the Dictator Game, where one person – the dictator – decides how to split an endowment between herself and another person – the recipient – who has no say, people behave in a more strategic and selfish way when deciding in groups compared to individual decision-making. For example, Dana et al. (2007) find that in a situation where two dictators decide simultaneously and the selfish outcome is implemented only if both dictators agree on it, 65% of all dictators choose the selfish option, while only 26% of all dictators choose the selfish option when deciding alone. This observation is confirmed by Luhan et al. (2009). In their experiment 23.4% of a dictator's endowment is sent to the recipient team consisting of three subjects when the dictator decides alone but only 10.8% is sent to the recipients when the dictator acts as a members of a three-person team. A similar pattern is found by Panchanathan et al. (2013). In their experiment 27.8% of a dictator's endowment is sent to a recipient when the dictator decides alone but only 11.61% of the endowment is sent to a recipient in a two dictator condition and only 8.8% of the endowment in a three dictator condition.

Although experimental evidence shows that people behave more selfishly in shared decisions, we do not know much about the driving forces behind it. Falk and Szech (2013) and Bartling et al. (2015) presume that individuals behave more selfishly when deciding in groups as the pivotality for the final outcome is diffused. This diffusion lowers the individual decisiveness for the final outcome and makes it easier to choose the self-interested option.

Engl (2018) builds upon the idea of pivotality and distinguishes between an ex-post and an ex-ante causal responsibility. Engl calls an agent *ex-post causally responsible* if, given the choices of all other decision makers, that agent's action turned out to be pivotal for the implementation of an outcome. If, prior to a decision, there is no uncertainty about choices of other decision makers and no uncertainty about other factors which could affect the outcome, then causal responsibility should be the same ex-post and ex-ante. It could be, however, that prior to a decision agents face uncertainty about the other decision makers' choices. Depending on what the other decision makers choose, the own action might or might not be pivotal. Engl, hence, defines *ex-ante causal responsibility* as the expected level of ex-post causal responsibility at the time when a decision is being made. Ex-ante causal responsibility takes into account the uncertainty about the pivotality of one's own decision. Adding another player to a decision and giving that player the power to prevent an outcome may, hence, lower ex-ante causal responsibility for that outcome.

According to Battigalli and Dufwenberg (2007), humans might aim at reducing the feeling of guilt caused by a decision. Building on this idea, Rothenhäusler et al. (2013) conclude that group-decisions allow to share the guilt for an individual decision and thus makes it easier to choose a selfish option in a group. In a similar spirit, Inderst et al. (2017) study an experiment where the causal attribution of guilt shifts from an advisor to a customer. Inderst et al. predict this shift by a model of shared guilt.

There are also concepts in social psychology explaining more selfish decision-making in groups than in individual decision situations. Darley and Latané (1968) propose the concept of *diffusion of responsibility*: selfish decisions in groups are caused by the possibility to share the responsibility for the outcome among group members. Several studies in social psychology confirm this idea. In a study by Forsyth et al. (2002) participants were asked to allocate 100 responsibility points among the members of the group (group size either 2, 4, 6, or 8 participants) after a group task was performed. The personal perceived responsibility for the group outcome was significantly lower the bigger the group. Freeman et al. (1975) study tipping behavior in restaurants. They show that people in groups tip on average less than individuals. Freeman et al. explain this finding with the diffused responsibility for tipping. Further possible mechanisms driving selfish decision-making in groups are suggested by research on the so-called *interindividual-intergroup discontinuity effect* by Insko et al. (1990), an effect that describes the tendency of individuals to be more competitive and less cooperative in groups than in one-on-one relations. According to this research, there are four moderators promoting selfish decisions in groups. First, the *social-support-for-shared-self-interest hypothesis* claims that group members can perceive active support for a self-interested choice by other group members. Second, the *identifiability hypothesis* proposes that deciding in groups provides a shield of anonymity that could also drive selfish decision-making. Third, according to the *ingroup-favoring norm*, decision makers could perceive some pressure to first benefit the own group before taking into account the interests of others. Finally, the *altruistic-rationalization hypothesis* suggests that deciding in a group enables individuals to justify their selfish behavior by arguing that the other group members will also benefit from it. According to a meta-study of 48 experiments on the *interindividual-intergroup discontinuity effect* by Wildschut et al. (2003) intergroup interactions are indeed in general more competitive than interindividual interactions.

To sum up, more selfish decision-making in groups seems to be driven by the diffused pivotality for the decision, a lower level of perceived responsibility and guilt for the outcome, the increased anonymity of the decision and the feeling that a selfish decision also favors the group and is supported or even demanded by the members of the group.

2.2. Perception of and behavior towards computers

A number of studies find that people treat computers in much the same way they treat people. For instance, Katagiri et al. (2001) show that people apply social norms from their own culture to a computer. Reeves and Nass (2003) found that people are as polite to computers as they are to humans in laboratory experiments. Nass et al. (1994) shows that people seem to use social rules in addressing computer behavior. Nass and Moon (2000) observe that people ascribe human-like attributes to computers. In a laboratory experiment by Nass et al. (1996), where subjects were told to be interdependent with a computer affiliate, the computer was perceived just like a human teammate. Moon and Nass (1998) even observe that humans tend to blame a computer for failure and take the credit for success when they feel dissimilar to it while blaming themselves for failure and crediting the computer for success when they feel similar to it. Other studies find that computers are held at least partly responsible for actions. Friedman (1995) reports in an interview on computer agency and moral responsibility for computer errors that 83% of the computer science major students attributed aspects of agency such as decision-making and/or intention to the computer, 21% of the students even held the computer moral responsible for wrongdoing. Moon (2003) shows that the self-serving tendency for the attribution of responsibility to a computer in a purchase decision experiment mitigates when the subjects have a history of intimate self-disclosure with a computer. In short, subjects' willingness to assign more responsibility to a computer for a positive outcome and less responsibility to the computer in a negative outcome increased, when the subjects shared some private information with the computer before the computer-aided purchase decision.

Although humans seem to treat computers and humans often in a similar way, differences remain: de Melo and Gratch (2015) find that recipients in a Dictator Game expect more money from a machine than from another human, and that proposers in an Ultimatum Game offer more money to a human recipient than to an artificial counterpart. de Melo and Gratch also show that people are more likely to perceive guilt when interacting with a human counterpart than when interacting with machines. Gogoll and Uhl (2016) find that people seem to dislike the usage of computers in situations where decisions affect a third party. In their experiment, people could delegate a decision in a trust game either to a human or to a computer algorithm that exactly resembles the human behavior in a previous trust game. Gogoll and Uhl observe that only 26.52% of all subjects delegate their decision to the computer instead of to a human. Gogoll and Uhl also allowed impartial observers to reward or to punish actors depending on their delegation decision. They find that, independent of the outcome, impartial observers reward delegations to a human more than delegation to a computer.

Consequently, especially in domains in which fundamental human properties such as moral considerations and ethical norms are of importance, findings from human-human interactions cannot necessarily be directly transferred to human-computer interactions. Although research in economics and social psychology analyses shared decision-making between humans extensively there seems to be a gap when it comes to shared decision-making with artificial systems such as computers.

3. Experimental design

We implemented an experiment with the following elements: (i) a binary Dictator Game in which participants were able to choose between an equal and an unequal split, (ii) a questionnaire to measure the perceived responsibility and guilt, and (iii) a manipulation check in which participants were confronted with a hypothetical decision situation. The decision in the binary Dictator Game was made either by a single human dictator, by two, so to speak multiple, human dictators, or by a computer together with a human dictator.

3.1. General procedures

In each experimental session, the following procedure was used: upon arrival at the laboratory, participants were randomly seated and randomly assigned a role (Player X, Player Z, and, depending on the treatment, Player Y). All participants were informed that they would be playing a game with one or two other participants in the room and that the matching would be random and anonymous. They were also told that all members of all groups would be paid according to the choices made in that group. Payoffs were explained using a generic payoff table. A short quiz ensured that the task and the payoff representation was understood. After the quiz, the actual payoffs were shown to participants together with any other relevant information for the treatment.

All treatments were one-shot dictator games with a binary choice between an equal and an unequal (socially inefficient) wealth allocation. After making a choice and before being informed about the final outcome, subjects answered a questionnaire to determine their perceived level of responsibility and guilt. Each participant was paid in private at the end of the experiment. All experimental stimuli, as well as instructions, were presented through a computer interface. We framed the game as neutrally as possible, avoiding any loaded terms. Payoffs were displayed in Experimental Currency Units (ECU's) with an exchange rate from 1 ECU equals 2 Euro. The entire experiment was computerized using z-Tree (Fischbacher, 2007). All subjects were recruited via ORSEE (Greiner, 2004).

3.2. Treatments

We had three different treatments in total. One treatment, the so-called *Single Dictator Treatment* or *SDT*, involved two players, one dictator and a recipient. Two more treatments involved three players, two dictators and one recipient. In one of these treatments, the so called *Multiple Dictator Treatment* or *MDT*, all players were humans. In the other treatment, the so-called *Computer Dictator Treatment* or *CDT*, the decision of one of the dictators was not made by him/herself but instead of by a computer. To compare the three different treatments we used a between-subject design.

3.2.1. Single dictator treatment (SDT)

Payoffs for the SDT are shown in the left part of Table 1. The dictator – Player X – had to decide between an unequal allocation (Option A) and an equal allocation (Option B). When the dictator chose Option A (Option B) then (s)he received a payoff of 6 ECU (5 ECU) and the recipient – Player Z – received a payoff of 1 ECU (5 ECU).

3.2.2. Multiple dictator treatment (MDT)

Payoffs for the MDT are shown in the right part of Table 1. Dictators – Player X and Player Y – both made a choice that determined the payoff for both dictators and the recipient. The

SDT:

Player X's choices		
	A	X:6, Y:–, Z:1
	B	X:5, Y:–, Z:5

MDT and CDT:

Player X's choices		Player Y's choices: **A**	Player Y's choices: **B**
	A	X:6, Y:6, Z:1	X:5, Y:5, Z:5
	B	X:5, Y:5, Z:5	X:5, Y:5, Z:5

Table 1: Payoffs in the Binary Dictator Games.

unequal payoff was only implemented if both dictators chose Option A. In all other cases the equal allocation was implemented. For example, if both dictators chose Option A then both dictators received a payoff of 6 ECUs while the recipient – Player Z – received a payoff of 1 ECU, however, if at least one of the two dictators chose Option B then the dictators as well as the recipient received a payoff of 5 ECU.

When choosing payoffs for the MDT we have to keep in mind that there are only two players in the SDT, but three players in the MDT (and in the CDT below). How should we make payoffs in the different treatments comparable? Should we keep individual payoffs constant (thus having more money on the table in MDT and CDT) or should we keep total payoffs constant (thus having smaller individual incentives in MDT and CDT)? Here we follow the literature (e.g. Dana et al., 2007) and keep individual incentives constant between all treatments. We are aware that, as a result, relative efficiency losses as unequal allocations are not constant between the treatments.

3.2.3. Computer dictator treatment (CDT)

The CDT was identical to the MDT with one exception: One of the two dictators – Player Y – acted as a so-called "passive dictator". While still receiving payoffs for Player Y as given in Table 1, the passive dictator had no influence on the choice as a computer made the choice. The frequency with which the computer chose options A or B followed the frequency of choices in an earlier MDT. Participants in the CDT were informed that frequencies were the same as in an earlier MDT treatment. Hence, all Players X in the CDT had the same beliefs (and the same uncertainty) about the other players' behavior as in the MDT. Furthermore, since payoff rules for Player Y in CDT were the same as in MDT, social concerns should not differ between CDT and MDT.

3.3. Measurement of perceived responsibility and guilt

After the dictators made their choices but before participants were informed about the final outcome and payoff, dictators completed a questionnaire. They had to state their perceived personal responsibility for the outcome. They also described their feeling of guilt if the unequal payoff allocation were to be implemented.[15] Dictator(s) were also asked to state their

[15] The wording of the questionnaire is provided in Appendix A.1.2.

perceived responsibility for the payoff of the recipient, and, depending on the treatment, for the payoff of the co-dictator. Similar to Forsyth et al. (2002) the perceived and allocated responsibility was measured on a scale from 0 to 100 using a slider. We used these questions as a proxy for the perceived responsibility and guilt for the final outcome and the perceived responsibility for the other participants. Subjects could also explain why they had chosen a specific option. Furthermore, in MDT and CDT, dictators were asked to state what they expected the other human co-dictator, alternatively the computer, to choose and how responsible and guilty they would perceive the human co-dictator or the computer to be if the unequal payoff allocation was implemented.

Recipients and, depending on the treatment, passive dictators were asked how they would assess the responsibility and guilt felt by the dictators if the unequal payoff allocation was implemented. They were also asked about their expectation how the dictator(s) decide and could state why they expected the dictator(s) to choose a specific option.

In a manipulation we asked how participants (dictators, recipients and, if present, passive dictators) would evaluate the situation used in the other treatment. We also collected some demographic data. Data and methods are available online.[16]

4. Theoretical framework and behavioral hypotheses

A purely selfish participant would take into account neither the welfare of others nor situational circumstances. In particular, for a selfish participant it should not matter whether the decision was made alone, with another person or with a computer. Similarly, for a participant with fixed social preferences the type of interaction partner, human or computer, should not matter. However, we know that social preferences depend on the salience of the link between actions and consequences. Chen and Schonger (2013) as well as Haisley and Weber (2010) show that certainty or ambiguity of the outcome matters. Grossman and van der Weele (2013), Grossman (2014) and Matthey and Regner (2011) argue that social preferences are affected by the availability of excuses which allow individuals to justify a selfish behavior. These findings can be supported with the help of models of social image concerns (e.g., Andreoni and Bernheim, 2009; Bénabou and Tirole, 2006; Ellingsen and Johannesson, 2008; Grossman, 2015) and models on self-perception maintenance (e.g., Aronson, 2009; Beauvois and Joule, 1996; Bodner and Prelec, 2003; Konow, 2000; Mazar et al., 2008; Murnighan et al., 2001; Rabin, 1995). According to these models, individuals not only maximize their own output but also want to be perceived by others as kind and fair and want to see themselves in a positive light. However, if these two goals are at odds, choosing an option that maximizes own output causes an unpleasant tension for the individual that can only be reduced by lowering the perceived conflict of interest between the two goals.[17] Therefore, as research in social psychology has shown, people seem to act selectively and in a self-serving way when determining whether a self-interested behavior will have a positive or negative impact on

[16] `https://www.kirchkamp.de/research/shareMachine.html` or `http://christina-strobel.de/share_responsibility.html`.

[17] The unpleasant tension (or in a more formal speech "disutility") is often described as nothing else than the feeling of guilt (e.g., Berndsen and Manstead, 2007; de Hooge et al., 2011; Stice, 1992).

their self-concept or social image and use situational excuses, if available, to justify their decision (e.g., Rabin, 1995; Haidt and Kesebir, 2010). In this way, individuals can blame selfish actions on the context in which they were made rather than on themselves, thus preserving a comfortable self-image.

Falk and Szech (2013) as well as Bartling et al. (2015) argue that in a situation where a decision is shared, decision makers are only responsible for a fraction of that decision as the pivotality for the decision is diffused. This diffusion provides an excuse to reduce responsibility for the final outcome. In short, sharing a decision with another human reduces the perceived negative consequences for the self- and social-image which makes it easier to choose a self-serving option.

According to the causal responsibility theory by Engl (2018), it is ex-ante, not ex-post causal responsibility that matters here. If a decision is shared each agent faces uncertainty about the behavior of the other agent. This uncertainty about behavior implies uncertainty about the final pivotality of the own decision. Uncertainty about own pivotality implies a lower perceived causal responsibility. More precisely, dictators can not be seen as less responsible for the outcome when choosing the fair option in the MDT than in the SDT as each dictator can ensure the implementation of the fair option independent of the other dictator. However, when choosing the unfair option dictators are less responsible for a final unfair outcome in the MDT than in the SDT as the final outcome is not only determined by one's individual choice but also by another person's decision. Thus, the ex-ante perceived causal responsibility in joint decision situations is lower compared to situations where a decision is not shared as the impact of one's decision has not yet been set.

Responsibility is also closely linked to interpersonal guilt. According to research in psychology, for example by Baumeister et al. (1994), Gilbert (1998), Tangney (1995), Tangney and Dearing (2003) and Hauge (2016), guilt is intimately tied to a person's recognition of being responsible for some wrongdoing. Whereby, guilty feelings may rise when a person feels responsible for or refraining from an action resulting in the infliction of harm or damage to another person or for failing to meet another persons expectations or moral standards.[18]

The theoretical arguments are supported by experimental evidence. Berndsen and Manstead (2007) show that the less responsible an individual feels, the less guilty the individual feels for making a selfish decision.

In our experiment, Option B leads to an equal payoff for all participants. However, if all decision makers choose Option A, the recipient receives much less than the dictator(s). Option A, hence, might cause more harm to the social and self-image than Option B. Dictators who value a positive perception by others and themselves more than their monetary gain will have a preference for Option B. Dictators who value mainly the monetary gain will prefer Option A.

In the SDT, the final payoffs only depend on the choice of a single dictator. The game offers no situational excuse to reduce the negative impact on the self- and social image caused by a selfish decision. Sharing a decision with another decision maker, however, provides the

[18]While the emotion of guilt can be evoked by various causes, the economic literature on guilt aversion mainly focuses on guilt caused by failing to meet others' expectations (e.g. Battigalli and Dufwenberg, 2007; Charness and Dufwenberg, 2006; Balafoutas and Sutter, 2017; Bellemare et al., 2018).

possibility to share the responsibility for the decision. In addition, deciding together with another dictator also creates room for the interpretation of a selfish choice as also beneficial for the other decider and creates complicity. This allows the dictator to attribute a selfish decision to the situation or circumstance rather than to his/her self-concept.[19] Furthermore, Bartling and Fischbacher (2012) show that people perceive others to be less responsible and blame them less for a negative outcome the less their decision influences the final decision. Thus, based on the shared responsibility for the decision as well as the anticipated lower blame by others for a selfish decision dictators in the MDT should be expected to experience less blame and thus to feel less guilty for a selfish choice compared to dictators in the SDT.

Hence, we expect that dictators in the MDT perceive themselves to be less responsible for the final outcome (Hypothesis 1.i) and to feel less guilty for a selfish decision (Hypothesis 2.i) than dictators in the SDT. As a result we expect more selfish decisions in the MDT than in the SDT (Hypothesis 3.i).

Turning to the CDT we must ask whether computer dictators are as responsible as human dictators. Can computers be in the same way responsible for an action? The concept of causal responsibility by Engl (2018) does not distinguish between human and computer opponents. In our experiment choices of computers follow the frequencies of choices of humans. Hence, both ex-ante and ex-post causal responsibility should be identical in MDT and in CDT.

In the literature we find the following three conditions required to be held responsible: First, an agent needs to have action power. Action power requires a causal relationship between own actions and the outcome (e.g., Lipinski et al., 2002; May, 1992; Moore, 1999; Nissenbaum, 1994; Scheines, 2002). Second, the agent must be able to choose freely. Free choice includes the competence to act on the basis of own authentic thoughts and motivations as well as the capability to control one's behavior (e.g., Fischer, 1999; Johnson, 2006). Third, to be held responsible requires the ability to consider the possible consequences an action might cause (e.g., Bechel, 1985; Friedman and Kahn, 1992). Furthermore, some researchers argue that it is necessary to be capable of suffering or gaining from possible blame or praise and thus to be culpable for wrongdoing (e.g., Moor, 1985; Sherman, 1999; Wallace, 1994). These conditions would also have to be satisfied by a computer in order for it to be held responsible. While the causal responsibility of a computer for an outcome cannot be denied, a computer neither has a free will nor the freedom of action. A computer is also not able to consider possible consequences of its actions in the same way as a human. Furthermore, a computer is not capable of any kind of emotions. Hence, a computer does not fulfill several of the conditions under which one could hold the computer responsible to the same extent as a human.[20] Research in machine and roboter ethics attributes only operational responsibility to the most advanced machines today but denies any higher form of (moral) responsibility as today's machines still have a relatively low level of own autonomy and ethical sensitivity (e.g., Allen et al., 2000; DeBaets, 2014; Dennett, 1997; Sullins, 2006).

Based on these considerations, the responsibility for a selfish outcome cannot be shared

[19]However, as either dictator can independently implemented the equal outcome by choosing Option B the addition of a second dictator does not impede subjects from ensuring a fair outcome if they prefer it.

[20]For the discussion on the responsibility of computers see Bechel (1985), Friedman and Kahn (1992), Snapper (1985), and, more recently, Asaro (2011), Floridi and Sanders (2004), Johnson and Powers (2005), Sparrow (2007), and Stahl (2006).

with a computer to the same extent as it can with a human. Thus, upholding a positive self- and social image while deciding selfishly together with a computer should not be as easy as when deciding with another human. The same holds true for the perceived guilt for an unfair outcome. Others cannot blame a computer for causing damage to a person in the same way as a human can be blamed. Hence, the guilt for harming another person cannot be shared with a computer to the same extent as with another human.[21]

For these reasons, we expect dictators to perceive more personal responsibility for the final outcome in the CDT than in the MDT (Hypothesis 1.ii). We also expect them to perceive more guilt when choosing the unfair option (Hypothesis 2.ii) in the CDT than in the MDT.

In addition, as selfish decision-making is influenced by the individual's perception of being responsible or feeling guilty for a decision, significantly more people should choose the selfish option if they are deciding with another human (MDT) than when deciding with a computer (CDT) (Hypothesis 3.ii).

Hypothesis 1 (responsibility) *In MDT participants attribute less responsibility to an individual dictator for the outcome resulting from choosing the selfish option than*

(i) in SDT, or

(ii) in CDT.

Hypothesis 2 (guilt) *In MDT participants attribute less guilt to an individual dictator for the outcome resulting from choosing the selfish option than*

(i) in SDT, or

(ii) in CDT.

Hypothesis 3 (selfishness) *In MDT the selfish option is chosen more frequently than*

(i) in SDT, or

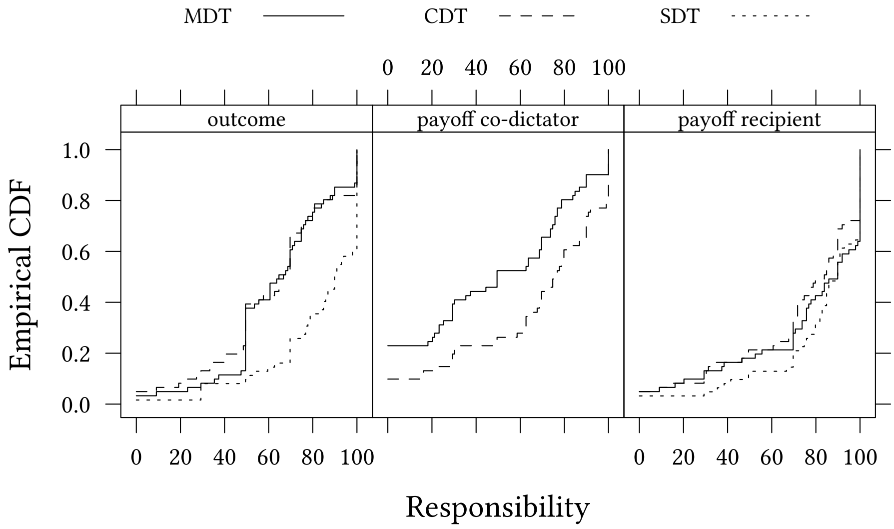
(ii) in CDT.

5. Results

All sessions were run in July, October and November 2016 at the Friedrich Schiller Universität Jena. Three treatments were conducted with a total of 399 subjects (65.2% female).[22] Most of our subjects were students with an average age of 25 years. Participants earned on average

[21]The beliefs about the choice of the computer or human co-dictator should also not differ as the frequency with which the computer chose options A or B followed the frequency of choices in an earlier MDT. Hence, the dictator's choice cannot be influenced by different beliefs about the frequency of unequal choices by humans or the computer.

[22]In total 124 subjects (62.9% female) participated in the SDT, 92 subjects (68.5% female) in the MDT and 183 subjects (65% female) in the CDT. Thus, we have almost the same number of actively deciding dictators in each treatment.



"Outcome" is Question 9 from Appendix A.1.2, "payoff co-dictator" is Question 7 from Appendix A.1.2, "payoff recipient" is from Question 6 from Appendix A.1.2.

Figure 1: Dictators' responsibility.

€9.43. We use a between-subject design, hence, the data for all statistical tests is independent for the different treatments.

We first analyze how the perceived responsibility for the final outcome as well as the feeling of guilt for a self-serving decision varied between the treatments before presenting the findings regarding the choices made by the dictators.

5.1. Hypothesis 1: responsibility

To assess perceived responsibility for a selfish decision we ask dictators to state their perceived level of responsibility for three different items: for the final outcome, for the recipient's payoff, and (in treatments MDT and CDT) for their co-dictator's payoff.[23] For all questions the level of responsibility was measured by a continuous scale from *"Not responsible at all"* (0) to *"Very responsible"* (100).

Figure 1 shows the distribution of personal responsibility for the three measures: outcome, payoff of the co-dictator, and payoff of the recipient. Figure 1 seems to confirm Hypothesis 1.i. According to this hypothesis, responsibility should be smaller in MDT than in SDT. Indeed, this seems to be the case for all three measures.

We find weaker support for Hypothesis 1.ii. According to this hypothesis, responsibility should be smaller in MDT than in CDT. This is clearly the case for responsibility for *payoff of co-dictator*. For the other two measures, however, the figure shows no clear difference between MDT and CDT.

[23]For the exact wording of the question for outcome see Question 9 from Appendix A.1.2. For the exact wording of the question for the recipient's payoff and the co-dictator's payoff see Questions 6 and 7 from Appendix A.1.2.

responsibility for...	SDT–MDT (Hyp. 1.i)		CDT–MDT (Hyp. 1.ii)	
outcome	$\Delta = 14.05$	CI=[7.627, 20.47] ($p = 0.0000$)	$\Delta = -2.35$	CI=[−8.855, 4.155] ($p = 0.4771$)
payoff co-dictator			$\Delta = 18.7$	CI=[6.574, 30.82] ($p = 0.0028$)
payoff recipient	$\Delta = 5.544$	CI=[−4.033, 15.12] ($p = 0.2538$)	$\Delta = -3.097$	CI=[−13.66, 7.466] ($p = 0.5627$)

The table shows differences between treatments ($\Delta = \ldots$), confidence intervals for this difference (CI=[...]), and p-values for a two sided test whether this difference could be zero. Each line shows the result for one measure: responsibility for outcome, responsibility for the co-dictator's payoff, responsibility for the recipient's payoff. Required effect sizes to reach significance are shown in Table 9 in Appendix A.12.

Table 2: Treatment difference in the dictator's responsibility.

Table 2 provides confidence intervals and p-values for treatment differences between the three measures. According to Hypothesis 1.i the difference in responsibility between SDT and MDT should be positive. Indeed, both the *outcome* measure and the *payoff recipient* measures are positive, however, only the *outcome* measure significantly so.[24]

According to Hypothesis 1.ii the difference in responsibility between CDT and MDT should be positive. We do observe a significant positive difference for the *payoff co-dictator* measure. However, we find insignificant negative differences for the other two measures.

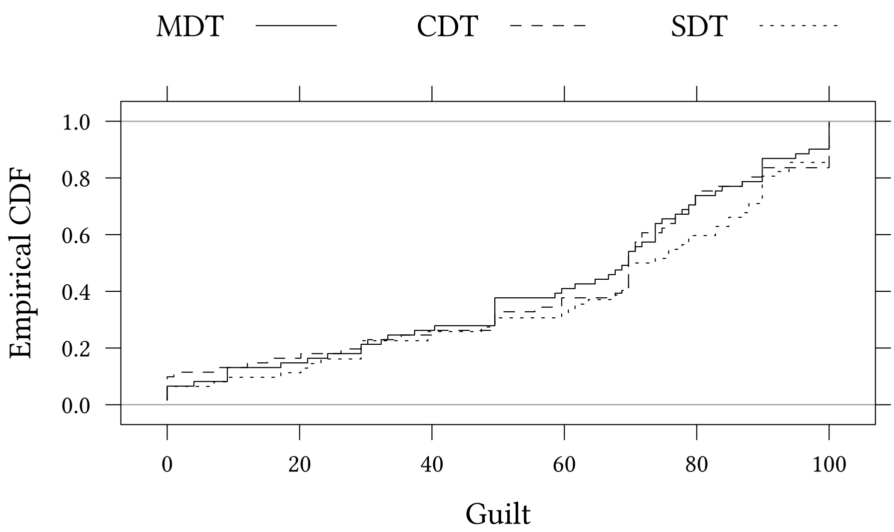
Further analysis showing how responsible the dictators perceived their co-dictators to be can be found in Appendix A.4 (see Table 6 and Figure 8). It shows that dictators perceive their fellow (human) dictator in the MDT as significantly more responsible than (computerized) dictators in the CDT. This difference is mainly driven by dictators who chose Option B.

5.2. Hypothesis 2: guilt

In all treatments dictators were asked to state their perceived guilt in case Option A was implemented. The level of guilt was measured by a continuous scale from *"not guilty"* (0) to *"totally guilty"* (100). Figure 2 shows the distribution of guilt. According to Hypothesis 2.i, we expect dictators to feel less guilty for an unequal payoff in the MDT than in the SDT. Furthermore, according to Hypothesis 2.ii we expect a lower level of guilt in MDT than in CDT. Table 3 provides confidence intervals and p-values for treatment differences. According to Hypothesis 2.i the difference in guilt between SDT and MDT should be positive. According to Hypothesis 2.ii the difference in guilt between CDT and MDT should be positive. Indeed, both differences are positive, however, not significantly so. Thus, neither Hypothesis 2.i nor Hypothesis 2.ii can be confirmed for dictators. The level of guilt felt by dictators is not significantly affected by the treatment, whether dictators decide on their own, together with a computer or with another human.

In Appendix A.10 we provide additional information about the guilt that recipients and

[24] Since in the SDT treatment there is no other dictator, we do not observe responsibility for the co-dictator's payoff.



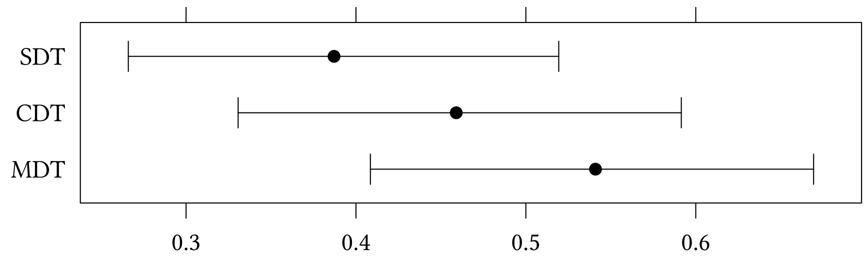
For the question see Question 8 from Appendix A.1.2.

Figure 2: Dictators' perceived guilt.

SDT–MDT		CDT–MDT	
$\Delta = 4.982$	CI=$[-6.093, 16.06]$ ($p = 0.3749$)	$\Delta = 0.9439$	CI=$[-10.35, 12.23]$ ($p = 0.8688$)

The table shows differences between treatments ($\Delta = \ldots$), confidence intervals for this difference (CI=[…]), and p-values for a two sided test whether this difference could be zero. Each line shows the result for one measure: responsibility for outcome, responsibility for the passive dictator's payoff, responsibility for the recipient's payoff. Required effect sizes to reach significance are shown in Table 10 in Appendix A.12.

Table 3: Treatment difference in guilt.



The graph shows 95%-confidence intervals around the observed frequency. For the question see Figure 4 in Appendix A.1.

Figure 3: Relative frequency of selfish choices by treatments.

passive dictators attribute to dictators (see Figure 18). We find that recipients in the MDT and in the CDT did not expect the dictators to feel significantly more guilty than recipients when choosing Option A.

5.3. Hypothesis 3: choices

Figure 3 presents, for each treatment, the relative frequency of self-interested choices made by dictators.[25] According to Hypothesis 3.i selfish choices should be more frequent in the MDT than in the SDT. Indeed, this is what we see in the figure. The difference is, however, not significant. According to Hypothesis 3.ii selfish choices should also be more frequent in the MDT than in the CDT. Again, this is what we see in the figure. Still, the difference is not significant.

6. Conclusion

The number of decisions made by human-computer teams has risen substantially in the past. Here, we study whether humans perceive a decision shared with a computer differently than a decision shared with another human. More specifically, we focus on the perceived personal responsibility and guilt for a selfish decision when a decision is shared with a computer instead of with another human.

Previous studies have established that humans behave more selfishly if they share responsibility with other humans. We do find a similar pattern in our experiment, even for human-computer interactions. When decision makers decide on their own, the number of selfish choices is rather small. When the decision is shared with a computer the number of selfish choices increases. The frequency of selfish choices is highest when the decision is shared with another human. However, these differences are not very large and, in our study, not

[25]For the binary Dictator Game interface shown to the dictators and to the recipients see Appendix A.1.1.

significant.[26] We also measure perceived responsibility for the final outcome, the recipient's payoff and the co-dictator's payoff. In line with our hypotheses, we find that responsibility for the outcome is perceived significantly stronger when a decision is not shared at all than when it is shared with a human. Also in line with our hypotheses, responsibility for the co-dictator's payoff is perceived stronger when the decision is shared with a computer than when the decision is shared with a human. Guilt, however seems to be perceived rather similarly – regardless whether a decision is shared with a human, with a machine or whether it is not shared at all. Participants did not perceive more guilt when deciding on their own or together with the computer than when deciding together with another human.

In our experiment we use a very small manipulation. The way the computers decided was fully transparent and could be easily linked to human choices. In the experiment the advantage of such a transparent design is that we can clearly communicate to participants what computers do. Sharing a choice with a computer in our experiment is as foreseeable as sharing a choice with with a human. We did, on purpose, not model the unpredictability of a complex computerized choice. This would be a next step which we have to leave to future research.

For the future, an open discussion of hybrid-decision situations would be desirable. It might not only be important to address the technical question of what we can achieve by using artificial decision making systems such as computer but also how humans perceive them in different situations and how this influences human decision-making.

References

Allen, C., Varner, G., and Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(3):251–261.

Andreoni, J. and Bernheim, B. D. (2009). Social image and the 50-50 norm: A theoretical and experimental analysis of audience effects. *Econometrica*, 77(5):1607–1636.

Andreoni, J. and Petrie, R. (2004). Public goods experiments without confidentiality: A glimpse into fund-raising. *Journal of Public Economics*, 88(7-8):1605–1623.

Aronson, E. (2009). The return of the repressed: Dissonance theory makes a comeback. *Psychological Inquiry*, 3(4):303–311.

Asaro, P. M. (2011). A body to kick, but still no soul to damn: Legal perspectives on robotics. In Lin, P., Abney, K., and Bekey, G. A., editors, *Robot Ethics: The Ethical and Social Implications of Robotics*, pages 169–186. MIT Press, Cambridge, MA.

[26] While the effect size for shared decision-making with another human was very large in the studies by Dana et al. (2007) ($n = 20$), Luhan et al. (2009) ($n = 30$) and Panchanathan et al. (2013) ($n = 44$), we found a medium sized effect ($n = 61$) when comparing decisions made by a single dictator to decisions made by a team of two human dictators.

Balafoutas, L. and Sutter, M. (2017). On the nature of guilt aversion: Insights from a new methodology in the dictator game. *Journal of Behavioral and Experimental Finance*, 13:9–15.

Bartling, B. and Fischbacher, U. (2012). Shifting the blame: On delegation and responsibility. *The Review of Economic Studies*, 79(1):67–87.

Bartling, B., Fischbacher, U., and Schudy, S. (2015). Pivotality and responsibility attribution in sequential voting. *Journal of Public Economics*, 128:133–139.

Battigalli, P. and Dufwenberg, M. (2007). Guilt in games. *American Economic Review*, 97(2):170–176.

Baumeister, R. F., Stillwell, A. M., and Heatherton, T. F. (1994). Guilt: an interpersonal approach. *Psychological Bulletin*, 115(2):243–267.

Beauvois, J.-L. and Joule, R. (1996). *A radical dissonance theory*. Taylor & Francis, London; Bristol, PA.

Bechel, W. (1985). Attributing responsibility to computer systems. *Metaphilosophy*, 16(4):296–306.

Bellemare, C., Sebald, A., and Suetens, S. (2018). Heterogeneous guilt sensitivities and incentive effects. *Experimental Economics*, 21(2):316–336.

Bénabou, R. and Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, 96(5):1652–1678.

Berndsen, M. and Manstead, A. S. R. (2007). On the relationship between responsibility and guilt: Antecedent appraisal or elaborated appraisal? *European Journal of Social Psychology*, 37(4):774–792.

Bland, J. and Nikiforakis, N. (2015). Coordination with third-party externalities. *European Economic Review*, 80:1–15.

Bodner, R. and Prelec, D. (2003). Self-signaling and diagnostic utility in everyday decision making. *The psychology of economic decisions*, 1:105–126.

Bornstein, G. and Yaniv, I. (1998). Individual and group behavior in the ultimatum game: Are groups more "rational" players? *Experimental Economics*, 1(1):101–108.

Botti, S. and McGill, A. L. (2006). When choosing is not deciding: The effect of perceived responsibility on satisfaction. *Journal of Consumer Research*, 33(2):211–219.

Burnette, J. L. and Forsyth, D. R. (2008). 'i didn't do it:' responsibility biases in open and closed groups. *Group Dynamics: Theory, Research, and Practice*, 12(3):210–222.

Charness, G. and Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, 74(6):1579–1601.

Chen, D. L. and Schonger, M. (2013). Social preferences or sacred values? Theory and evidence of deontological motivations. *Working Paper, ETH Zürich, Mimeo.*

Cooper, D. J. and Kagel, J. H. (2005). Are two heads better than one? Team versus individual play in signaling games. *American Economic Review*, 95(3):477–509.

Dana, J., Weber, R. A., and Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1):67–80.

Darley, J. and Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, 8(4, Pt.1):377–383.

de Hooge, I. E., Nelissen, R. M. A., Breugelmans, S. M., and Zeelenberg, M. (2011). What is moral about guilt? Acting "prosocially" at the disadvantage of others. *Journal of Personality and Social Psychology*, 100(3):462–473.

de Melo, C. M. and Gratch, J. (2015). People show envy, not guilt, when making decisions with machines. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 315–321. IEEE.

de Melo, C. M., Marsella, S., and Gratch, J. (2016). People do not feel guilty about exploiting machines. *ACM Transactions on Computer-Human Interaction*, 23(2):1–17.

DeBaets, A. M. (2014). Can a robot pursue the good? Exploring artificial moral agency. *Journal of Evolution and Technology*, 24:76–86.

Dennett, D. C. (1997). When HAL kills, who's to blame?: Computer ethics. *Rethinking responsibility in science and technology*, pages 203–214.

Ellingsen, T. and Johannesson, M. (2008). Pride and prejudice: The human side of incentive theory. *American Economic Review*, 98(3):990–1008.

Engel, C. (2011). Dictator games: A meta study. *Experimental Economics*, 14(4):583–610.

Engl, F. (2018). A theory of causal responsibility attribution. *University of Cologne Working paper 2932769.*

Falk, A. and Szech, N. (2013). Morals and markets. *Science*, 340(6133):707–711.

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

Fischer, J. M. (1999). Recent work on moral responsibility. *Ethics*, 110(1):93–139.

Fischer, P., Krueger, J. I., Greitemeyer, T., Vogrincic, C., Kastenmüller, A., Frey, D., Heene, M., Wicher, M., and Kainbacher, M. (2011). The bystander-effect: A meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin*, 137(4):517–537.

Floridi, L. and Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, 14(3):349–379.

Forsyth, D. R., Zyzniewski, L. E., and Giammanco, C. A. (2002). Responsibility diffusion in cooperative collectives. *Personality and Social Psychology Bulletin*, 28(1):54–65.

Freeman, S., Walker, M. R., Borden, R., and Latane, B. (1975). Diffusion of responsibility and restaurant tipping: Cheaper by the bunch. *Personality and Social Psychology Bulletin*, 1(4):584–587.

Friedman (1995). "It's the computer's fault"—Reasoning about computers as moral agents. In *Conference companion on Human factors in computing systems (CHI 95)*, pages 226–227. Association for Computing Machinery, New York, NY.

Friedman, B. and Kahn, P. H. (1992). Human agency and responsible computing: Implications for computer system design. *Journal of Systems and Software*, 17(1):7–14.

Gilbert, P. (1998). What is shame? some core issues and controversies. In Kremer, K. and Macho, V., editors, *Shame:Interpersonal Behavior, Psychopathology and Cultures*, pages 3–38. Oxford University Press, New York.

Gogoll, J. and Uhl, M. (2016). Automation and morals — eliciting folk intuitions. *TU München Peter Löscher-Stiftungslehrstuhl für Wirtschaftsethik Working Paper Series.*

Gosling, P., Denizeau, M., and Oberlé, D. (2006). Denial of responsibility: a new mode of dissonance reduction. *Journal of personality and social psychology*, 90(5):722–733.

Greiner, B. (2004). An online recruitment system for economic experiments. In Kremer, K. and Macho, V., editors, *Forschung und wissenschaftliches Rechnen*, pages 79–93. Göttingen.

Grossman, Z. (2014). Strategic ignorance and the robustness of social preferences. *Management Science*, 60(11):2659–2665.

Grossman, Z. (2015). Self-signaling and social-signaling in giving. *Journal of Economic Behavior & Organization*, 117:26–39.

Grossman, Z. and van der Weele, J. J. (2013). Self-image and willful ignorance in social decisions. *Forthcoming in the Journal of the European Economic Association.*

Haidt, J. and Kesebir, S. (2010). Morality. In Fiske, S. T., Gilbert, D. T., Lindzey, G., and Jongsma, Jr., A. E., editors, *Handbook of social psychology*. Wiley, Hoboken, N.J.

Haisley, E. C. and Weber, R. A. (2010). Self-serving interpretations of ambiguity in other-regarding behavior. *Games and Economic Behavior*, 68(2):614–625.

Hauge, K. E. (2016). Generosity and guilt: The role of beliefs and moral standards of others. *Journal of Economic Psychology*, 54:35–43.

Inderst, R., Khalmetski, K., and Ockenfels, A. (2017). Sharing guilt: How better access to information may backfire. *University of CologneWorking Paper Series in Economics*, (90).

Insko, C. A., Schopler, J., Hoyle, R. H., Dardis, G. J., and Graetz, K. A. (1990). Individual-group discontinuity as a function of fear and greed. *Journal of Personality and Social Psychology*, 58(1):68–79.

Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology*, 8(4):195–204.

Johnson, D. G. and Powers, T. M. (2005). Computer systems and responsibility: A normative look at technological complexity. *Ethics and Information Technology*, 7(2):99–107.

Katagiri, Y., Nass, C., Takeuchi, and Yugo (2001). Cross-cultural studies of the computers are social actors paradigma: The case of reciprocity. In Smith, M. J., Koubek, R. J., Salvendy, G., and Harris, D., editors, *Usability evaluation and interface design*, volume 1 of *Human factors and ergonomics*, pages 1558–1562. Lawrence Erlbaum, Mahwah, N.J. and London.

Kaushik, D., High, R., Clark, C. J., and LaGrange, C. A. (2010). Malfunction of the Da Vinci robotic system during robot-assisted laparoscopic prostatectomy: an international survey. *Journal of endourology*, 24(4):571–575.

Kocher, M. G. and Sutter, M. (2005). The decision maker matters: Individual versus group behaviour in experimental beauty-contest games. *The Economic Journal*, 115(500):200–223.

Kocher, M. G. and Sutter, M. (2007). Individual versus group behavior and the role of the decision making procedure in gift-exchange experiments. *Empirica*, 34(1):63–88.

Konow, J. (2000). Fair shares: Accountability and cognitive dissonance in allocation decisions. *American Economic Review*, 90(4):1072–1092.

Kugler, T., Bornstein, G., Kocher, M. G., and Sutter, M. (2007). Trust between individuals and groups: Groups are less trusting than individuals but just as trustworthy. *Journal of Economic Psychology*, 28(6):646–657.

Latané, B. and Nida, S. (1981). Ten years of research on group size and helping. *Psychological Bulletin*, 89(2):308–324.

Lipinski, T. A., Buchanan, E. A., and Britz, J. J. (2002). Sticks and stones and words that harm: Liability vs. responsibility, section 230 and defamatory speech in cyberspace. *Ethics and Information Technology*, 4(2):143–158.

Luhan, W., Kocher, M., and Sutter, M. (2009). Group polarization in the team dictator game reconsidered. *Experimental Economics*, 12(1):26–41.

Matthey, A. and Regner, T. (2011). Do i really want to know? A cognitive dissonance-based explanation of other-regarding behavior. *Games*, 2(4):114–135.

May, L. (1992). *Sharing responsibility.* University of Chicago Press, Chicago.

Mazar, N., Amir, O., and Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, 45(6):633–644.

McGlynn, R. P., Harding, D. J., and Cottle, J. L. (2009). Individual-group discontinuity in group-individual interactions: Does size matter? *Group Processes & Intergroup Relations*, 12(1):129–143.

Moon, Y. (2003). Don't blame the computer: When self-disclosure moderates the self-serving bias. *Journal of Consumer Psychology*, 13(1-2):125–137.

Moon, Y. and Nass, C. (1998). Are computers scapegoats? Attributions of responsibility in human–computer interaction. *International Journal of Human-Computer Studies*, 49(1):79–94.

Moor, J. H. (1985). Are there decisions computers should never make? In Johnson, D. G. and Snapper, J. W., editors, *Ethical issues in the use of computers*, pages 120–130. Wadsworth Publ. Co., Belmont, CA.

Moore, M. S. (1999). Causation and responsibility. *Social Philosophy and Policy*, 16(2):1–51.

Murnighan, J., Oesch, J. M., and Pillutla, M. (2001). Player types and self-impression management in dictatorship games: Two experiments. *Games and Economic Behavior*, 37(2):388–414.

Mynatt, C. and Sherman, S. J. (1975). Responsibility attribution in groups and individuals: A direct test of the diffusion of responsibility hypothesis. *Journal of Personality and Social Psychology*, 32(6):1111–1118.

Nass, C., Fogg, B. J., and Moon, Y. (1996). Can computers be teammates? *International Journal of Human-Computer Studies*, 45(6):669–678.

Nass, C. and Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1):81–103.

Nass, C., Steuer, J., and Tauber, E. R. (1994). Computers are social actors. In Adelson, B., Dumais, S., and Olson, J., editors, *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 72–78, New York. Association for Computing Machinery.

Nissenbaum, H. (1994). Computing and accountability. *Communications of the ACM*, 37(1):72–80.

Panchanathan, K., Frankenhuis, W. E., and Silk, J. B. (2013). The bystander effect in an n-person dictator game. *Organizational Behavior and Human Decision Processes*, 120(2):285–297.

R Development Core Team (2018). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rabin, M. (1995). Moral preferences, moral constraints, and self-serving biases. *Department of Economics UCB (unpublished manuscript).*

Reeves, B. and Nass, C. (2003). *The media equation: How people treat computers, television, and new media like real people and places.* CSLI Publ, Stanford CA, 1. paperback ed., 3. [print.] edition.

Reuben, E. and van Winden, F. (2010). Fairness perceptions and prosocial emotions in the power to take. *Journal of Economic Psychology*, 31(6):908–922.

Rockenbach, B., Sadrieh, A., and Mathauschek, B. (2007). Teams take the better risks. *Journal of Economic Behavior & Organization*, 63(3):412–422.

Rothenhäusler, D., Schweizer, N., and Szech, N. (2013). Institutions, shared guilt, and moral transgression. *Working Paper Series in Economics*, (47).

Savitsky, K., Van Boven, L., Epley, N., and Wight, W. M. (2005). The unpacking effect in allocations of responsibility for group tasks. *Journal of Experimental Social Psychology*, 41(5):447–457.

Scheines, R. (2002). Computation and causation. *Metaphilosophy*, 33(1/2):158–180.

Shademan, A., Decker, R. S., Opfermann, J. D., Leonard, S., Krieger, A., and Kim, P. C. W. (2016). Supervised autonomous robotic soft tissue surgery. *Science translational medicine*, 8(337):337ra64.

Sherman, N. (1999). Taking responsibility for our emotions. *Social Philosophy and Policy*, 16(02):294–323.

Snapper, J. W. (1985). Responsibility for computer–based errors. *Metaphilosophy*, 16(4):289–295.

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy*, 24(1):62–77.

Stahl, B. C. (2006). Responsible computers? A case for ascribing quasi-responsibility to computers independent of personhood or agency. *Ethics and Information Technology*, 8(4):205–213.

Stice, E. (1992). The similarities between cognitive dissonance and guilt: Confession as a relief of dissonance. *Current Psychology*, 11(1):69–77.

Sullins, J. P. (2006). When is a robot a moral agent? In Adelson, M. and Anderson, S., editors, *Machine Ethics*, pages 151–160, New York, NY. Association for Computing Machinery.

Sutter, M. (2005). Are four heads better than two? An experimental beauty-contest game with teams of different size. *Economics Letters*, 88(1):41–46.

Tangney, J. P. (1995). Shame and guilt in interpersonal relationships. In Tangney, J. P. and Fischer, K. W., editors, *Self-conscious emotions: Shame, guilt, embarrassment, and pride*, pages 114–139, New York. Guilford Press.

Tangney, J. P. and Dearing, R. L. (2003). *Shame and guilt.* Guilford Press.

Wallace, R. J. (1994). *Responsibility and the moral sentiments.* Harvard University Press, Cambridge, Mass.

Whyte, G. (1991). Diffusion of responsibility: Effects on the escalation tendency. *Journal of Applied Psychology*, 76(3):408–415.

Wildschut, T., Pinter, B., Vevea, J. L., Insko, C. A., and Schopler, J. (2003). Beyond the group mind: a quantitative review of the interindividual-intergroup discontinuity effect. *Psychological Bulletin*, 129(5):698–722.

A. Appendix for online publication

This section contains additional information on the interfaces and questions used in the treatments. We also present further analyses of data we collected in addition to the data used to test your hypotheses. Data and Methods can be found at `https://www.kirchkamp.de/research/shareMachine.html` or `http://christina-strobel.de/share_responsibility.html`.

A.1. Interfaces and questions

In this section the interfaces as well as the questions used in the experiment are presented.

A.1.1. Dictator game interface

In the MDT as well as in the CDT dictators used the interface sketched in Figure 4 to enter their decision. Recipients used the interface sketched in Figure 5 to enter their guess.

Figure 4: Dictator Game interface for dictators.

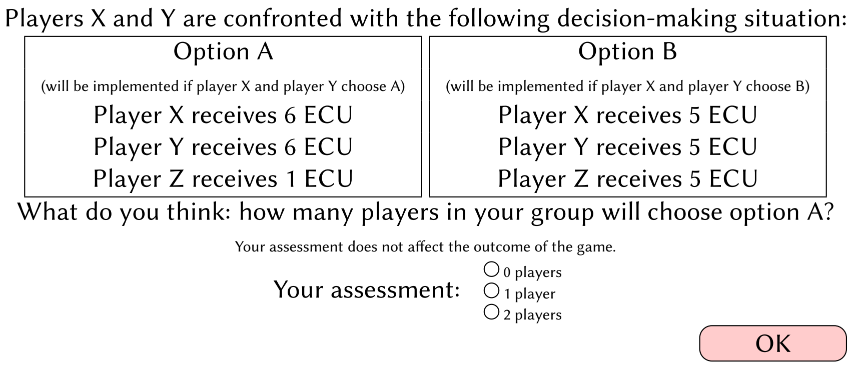


Figure 5: Dictator Game interface for recipients and passive dictators.

The interfaces for dictators and recipients were as similar as possible in all three treatments. Recipients were asked to guess dictators choices.

A.1.2. Questionnaire

All subjects were asked to complete a questionnaire. The questions were asked right after the decision and before the final outcome was announced. As an example, the questions used in the MDT for the subject in the role of Player X are presented below. The used answer method is presented in squared brackets. The questions asked in the CDT and in the SDT were very similar to the questions asked in the MDT. In the CDT, Player Y did not decide on his/her own, and the questions were changed accordingly. Except the first three questions, all questions were asked in the SDT. Dictators were asked directly, recipients and passive dictators were asked indirectly. For example, recipients and passive dictators were asked how responsible they perceive the dictator(s) to be for the recipients' or the passive dictators' payoff and how responsible they expect the dictator(s) to feel for the final outcome.

1. How would you have decided, if you had made the decision on your own? [Slider from *"Option A"* to *"Option B"*] (for an analysis of the answers given see Appendix A.5)
2. What is the likelihood that Player Y chooses Option A (Player X receives 6 ECU, Player Y receives 6 ECU, Player Z receives 1 ECU)? [Slider from *"Player Y always chooses A"* to *"Player Y always chooses B"*] (for an analysis of the answers given see Appendix A.6)
3. Did your expectation regarding the likelihood that Player Y would choose Option A (Player X receives 6 ECU, Player Y receives 6 ECU, Player Z receives 1 ECU) affect your decision? [*Radio buttons "YES"; "NO"*] (for an analysis of the answers given see Appendix A.2)
4. Why did you choose Option A (Player X receives 6 ECU, Player Y receives 6 ECU, Player Z receives 1 ECU)? [Open question with a maximum of 100 characters] / Why did you choose Option B (Player X receives 5 ECU, Player Y receives 5 ECU, Player Z receives 5 ECU)? [Open question with a maximum of 100 characters] (for the answers given see online dataset)
5. What could be additional reasons for choosing Option A (Player X receives 6 ECU, Player Y receives 6 ECU, Player Z receives 1 ECU)? [Open question with a maximum of 100 characters] (for the answers given see online dataset)
6. I feel responsible for the payoff of Player Z. [Slider from *"Very responsible"* to *"Not responsible at all"*] (for an analysis of the answers given see Section 5.1 and Appendix A.9)[27]
7. I feel responsible for the payoff of Player Y. [Slider from *"Very responsible"* to *"Not responsible at all"*] (for an analysis of the answers given see Section 5.1 and Appendix A.9)[28]

[27] Recipients and passive dictators were asked how responsible they perceive the dictator to be for the payoff of Player Z.

[28] Recipients and passive dictators were asked how responsible they perceive the dictator to be for the payoff of Player Y.

8. Option A will be implemented if you and the other player chose Option A. If this happens, Player X receives 6 ECU, Player Y receives 6 ECU and Player Z receives 1 ECU. Please adjust the slide control, so that it shows how guilty you would feel in this case? [Slider from *"I would feel very guilty"* to *"I would not feel guilty at all"*] (for an analysis of the answers given see Section 5.2 and Appendix A.10)[29]

9. Option A will be implemented if you and the other player chose Option A. In this case, Player X receives 6 ECU, Player Y receives 6 ECU and Player Z receives 1 ECU. Please adjust the slide control, so that it shows how you would perceive your responsibility as well as the responsibility of the other player in a scenario in which Option A is implemented. [Slider from *"I am fully responsible"* to *"I am not responsible"* and slider from *"My fellow player is fully responsible"* to *"My fellow player is not responsible"*] (for an analysis of the answers given see Section 5.1 and Appendix A.4 and A.8)[30]

A.2. Dictators' perceived influence by co-dictators choice

Dictators in the MDT as well as in the CDT were asked to state if their expectation regarding their co-dictators behavior had an influence on their own decision.[31] Dictators could either choose *"YES"* or *"NO"*. In the MDT 34.4% of the dictators and in the CDT 36.1% of the dictators stated that they took the expected decision of their co-dictator into account when making their own decision.

A.3. Dictators' responsibility for subsets of decisions

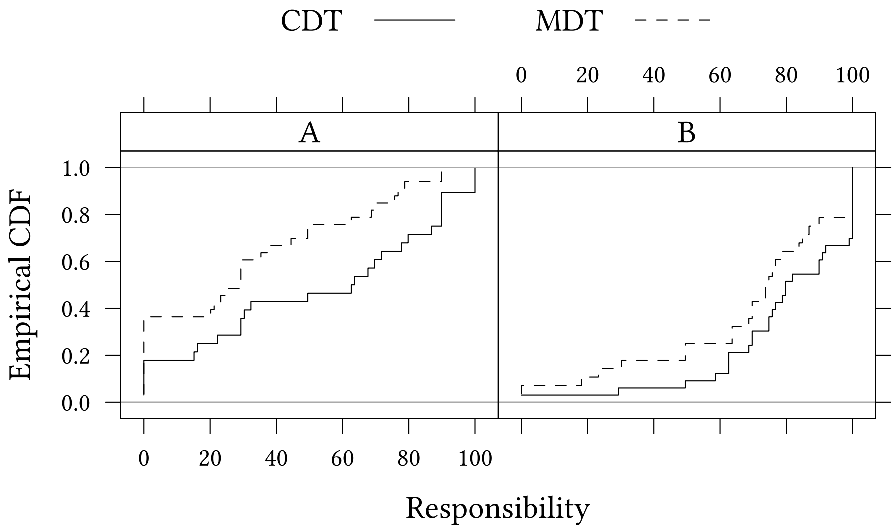
In the discussion of Table 2 we have seen that that dictators feel significantly more responsibility for the payoff of the co-dictator when they share a decision with a computer rather than with a human. Here we ask whether this effect is perhaps driven by a specific type of dictator – only those who choose Option A or only those who choose Option B. Figure 6 shows the responsibility for the co-dictator's payoff in relation to the dictator's own choice. We see that for both types of choices, Option A and Option B, dictators always feel more responsible in the CDT treatment than in the MDT treatment. This observation is confirmed by the data presented in Table 4. The mean perceived responsibility for the payoff of the co-dictator is always smaller in MDT than in CDT, regardless whether the dictator chose Option A or Option B.

In the discussion of Table 2 we have shown that there is only a small difference between the MDT and CDT treatments when it comes to responsibility for the outcome. We have seen that dictators actually feel a bit less responsible for the outcome in the CDT treatment than in the MDT treatment. Figure 7 and Table 5 illustrate that the difference between the CDT and MDT treatment does not depend on the dictator's choice. Regardless whether the

[29] Recipients and passive dictators were asked how guilty they expect the dictator to feel if Option A would be implemented.

[30] Recipients and passive dictators were asked how responsible they expect the dictator to feel if Option A would be implemented.

[31] For the exact wording of the question see Question 3 from Appendix A.1.2.



For the question see Question 7 from Appendix A.1.2.

Figure 6: Responsibility for the co-dictator's payoff depending on the dictator's own choice.

Payoff co-dictator	Option A	Option B
MDT	28.96	65.24
CDT	52.27	79.31

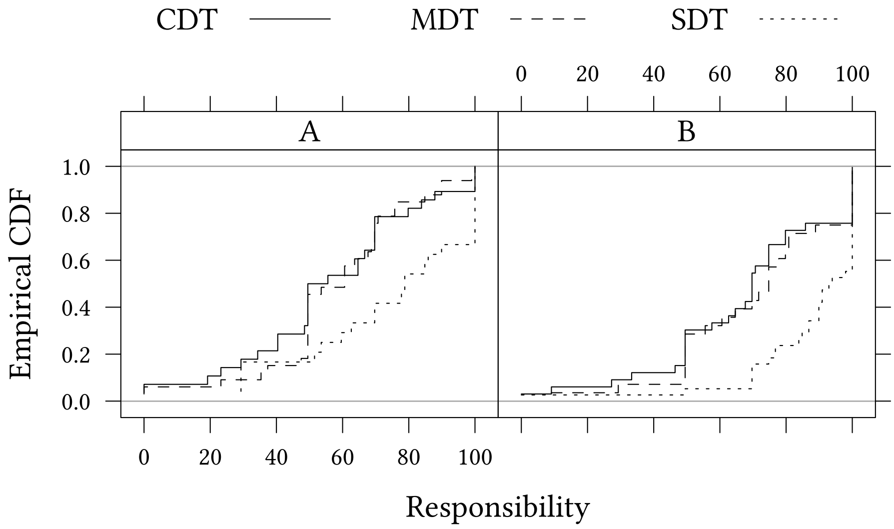
The table shows the mean of the perceived responsibility by the dictators for the payoff of the co-dictator separated for dictators who chose Option A and Option B.

Table 4: Treatment difference in the dictator's responsibility for the co-dictator's payoff by choice.

Unfair outcome	Option A	Option B
SDT	74.16	87.77
MDT	65.54	71.63
CDT	55.88	67.74

The table shows the mean of the perceived responsibility by the dictators for the outcome separated for dictators who chose Option A and Option B.

Table 5: Treatment difference in the dictator's perceived responsibility for an unfair outcome by choice.



For the question see Question 9 from Appendix A.1.2.

Figure 7: Responsibility for an unfair outcome (dictators).

CDT–MDT	
$\Delta = -15.1$	CI=$[-\infty, -9.349]$ ($p = 0.0000$)

Required effect sizes to reach significance are shown in Table 11 in Appendix A.12.

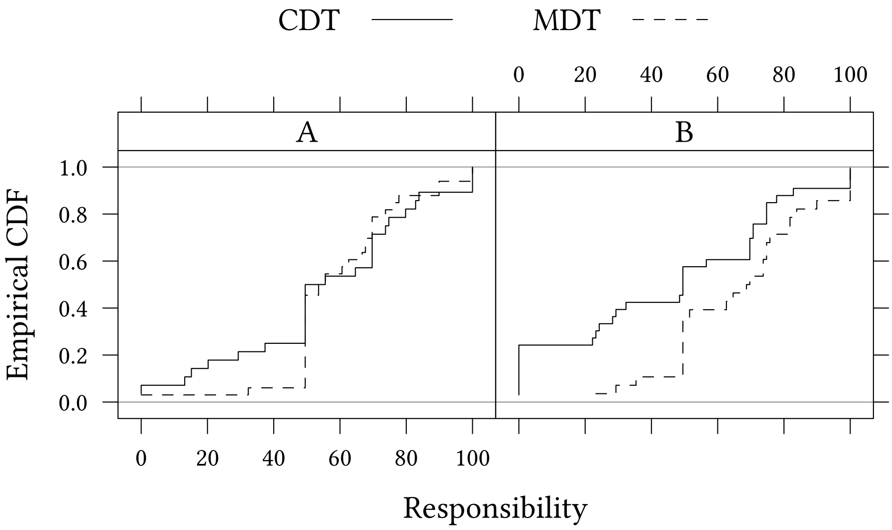
Table 6: Treatment difference between the personal responsibility of the computer in the CDT and the human dictator in the MDT as perceived by the dictators.

dictator chooses Option A or Option B, perceived responsibility is always a bit smaller in the CDT treatment.

A.4. Dictators' assigned responsibility to the co-dictator by choice

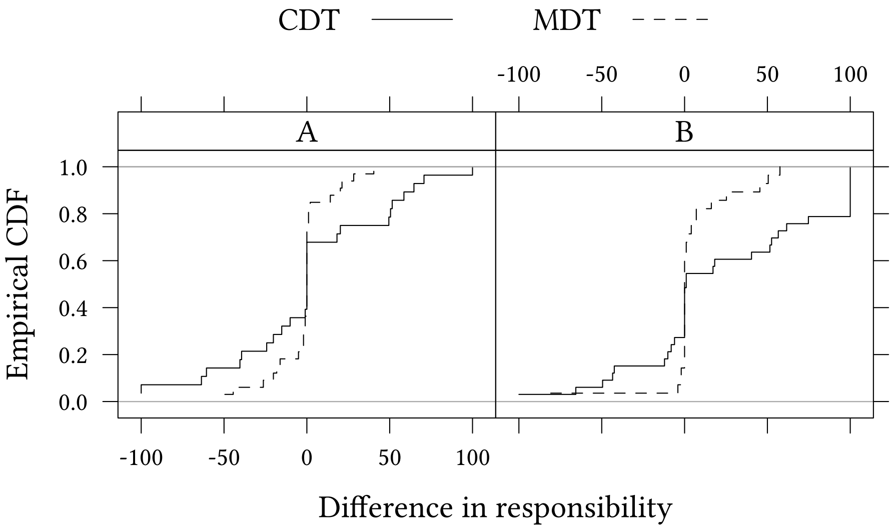
Dictators in the MDT and in the CDT had the possibility to state how responsible they perceive their co-dictator – either a human in the MDT or a computer in the CDT – to be for the final outcome in case Option A is implemented.[32] As Table 6 shows, dictators in the MDT perceived their human co-dictator, on average, to be significantly more responsible for the implementation of an unfair outcome than the dictators in the CDT perceived the computer to be. However, as Figure 8 shows, the level of responsibility assigned to the computer or human co-dictator differs between dictators who chose option A and dictators who chose option B. In case option A would be implemented dictators who had chosen option A, the unfair option, would allocate quite the same amount of responsibility to the computer in the CDT than to the human co-dictator in the MDT. Dictators who had chosen option B, however,

[32]For the exact wording of the question see Question 9 from Appendix A.1.2.



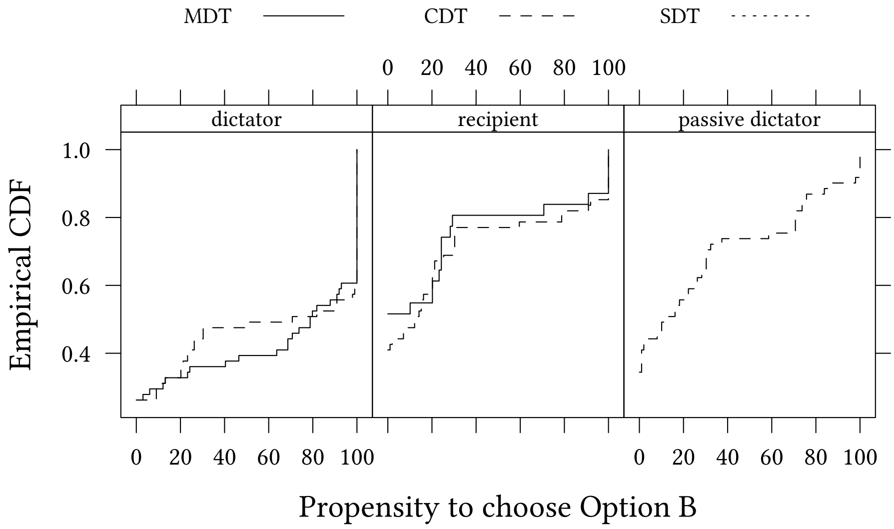
For the question see Question 9 from Appendix A.1.2.

Figure 8: Responsibility assigned to the computer or human co-dictator by dictators.



For the question see Question 9 from Appendix A.1.2.

Figure 9: Difference between dictators' personal responsibility and co-dictators' responsibility.



For the question see Question 1 from Appendix A.1.2. "Dictator" is the dictators' own assessment, "recipient" is how the recipients expect the dictators to decide as hypothetical single players, "passive dictator" is how the passive dictators expect the dictators to decide as hypothetical single players.

Figure 10: Dictators' choice as a hypothetical single player.

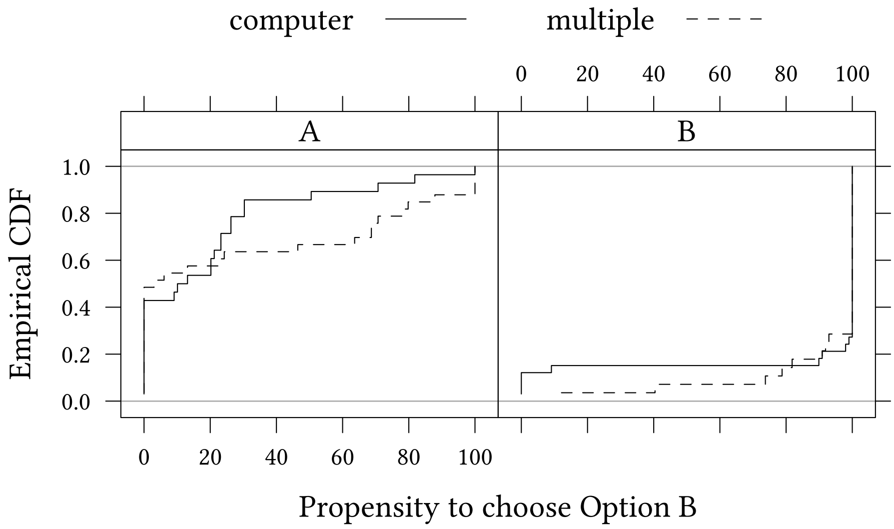
would perceive the computer in the CDT to be less responsible than a human co-dictator in the MDT.

By comparing the responsibility the dictators assigned to themselves with the responsibility the dictators attributed to their co-dictators, see Figure 9, it becomes clear that the difference is more dispersed in the CDT, where dictators decided together with a computer, than in the MDT, where dictators decided together with another human dictator. The means, however, are similar (p-value 0.1273). In summary, dictators assigned on average less responsibility to a computer in the CDT than to a human co-dictator in the MDT.

A.5. Hypothetical decision if dictators decide as single dictators

Dictators in the MDT as well as in the CDT were asked how they would have decided, if they would have had to decide on their own. Recipients in the MDT and CDT and passive dictators in the CDT were asked how they would have expected the dictator to decide, if they would have had to decide on their own.[33] Dictators as well as recipients were able to insert their assessment by using a continuous scale from *"Option A"* (0) to *"Option B"* (100). As the left part of Figure 10 shows, a large proportion of the actively deciding dictators in the CDT and in the MDT reported that they would have chosen Option B if they had been forced to decide alone. This was mainly driven by dictators who chose Option B (p-value 0.0000) (see Figure 11). As the middle part of Figure 10 shows, it become clear that recipients in the MDT as well as in the CDT expected the dictators to choose Option B less often if they would have

[33]For the exact wording of the question see Question 1 from Appendix A.1.2.



Question 1 from Appendix A.1.2.

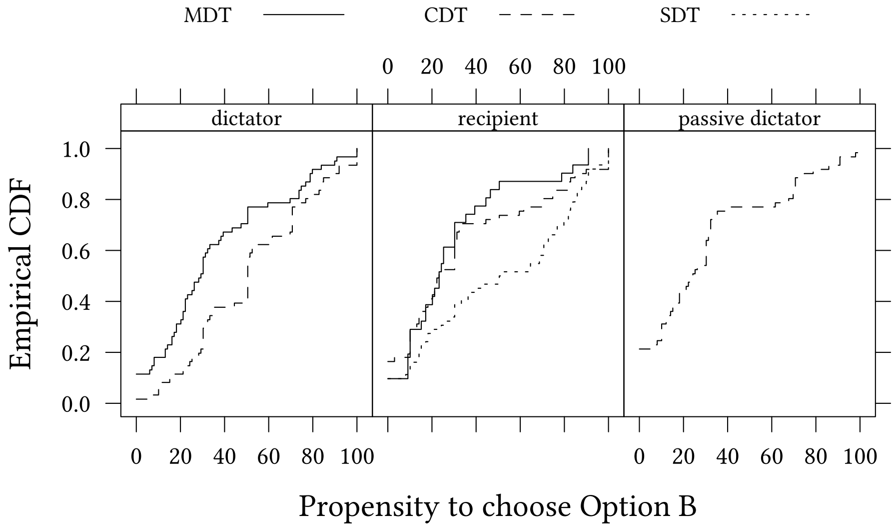
Figure 11: Dictators' choice as a hypothetical single player by choice.

to decide alone. As the right part of Figure 10 shows, the passive dictators in the CDT also expected the dictators to choose Option B less often where they deciding alone.

A.6. Expectation regarding the behavior of the human dictator(s) and the computer

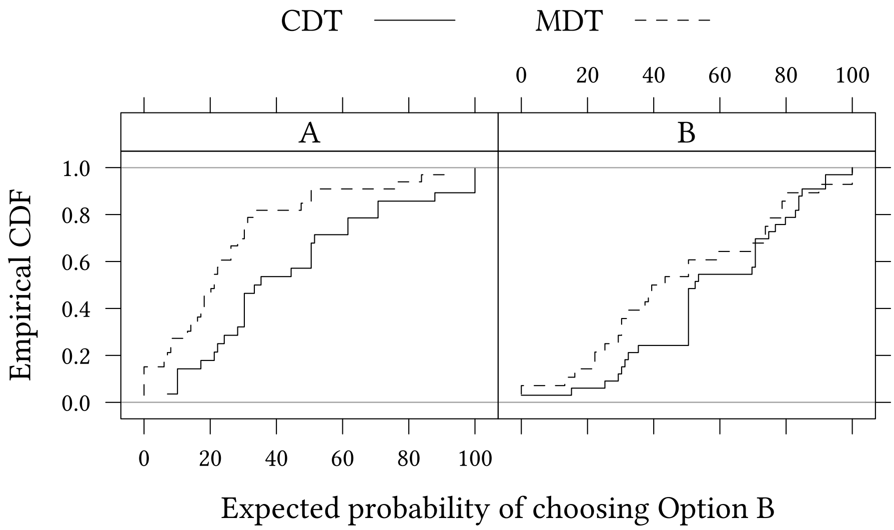
Dictators in the MDT as well as in the CDT were asked to state the likelihood that their co-dictator would choose Option A. Recipients in the MDT as well as in the CDT were asked to state the likelihood that the dictator as well as the co-dictator choose Option A.[34] Passive dictators in the CDT were asked to state what they expected the dictator to choose.[35] The expectation was measured by using a continuous scale from *"Player [Computer] always chooses Option A"* (0) to *"Player [Computer] always chooses always Option B"* (100). As the left part of Figure 12 shows, dictators in the CDT expected the computer to choose Option A on average significantly less often than dictators in the MDT expected their human co-dictator to choose Option A (p-value 0.0023). This was mainly driven by dictators in the MDT who had chosen Option B (p-value 0.0001) (see Figure 13). As the middle part of Figure 12 shows, recipients in the SDT expected dictators to be more likely to choose Option B than recipients in the MDT (p-value 0.0012). However, recipients in the MDT did not expect a higher likelyhood of selfish choices by dictators than recipients in the CDT (p-value 0.4382). As the right part of Figure 12 shows, passive dictators in the CDT expected the dictator to be more likely to choose Option A than Option B.

[34]In the MDT the co-dictator was another human, in the CDT the co-dictator was a computer.
[35]For the exact wording of the question see Question 2 from Appendix A.1.2.



For the question see Question 2 from Appendix A.1.2. “Dictator” is the dictator’s expectation about the co-dictator’s choice, “recipient” is what the recipients expect the dictators to choose, “passive dictator” is what the passive dictators expect the dictators to choose.

Figure 12: Expected co-dictator’s choice.



For the question see Question 2 from Appendix A.1.2.

Figure 13: Dictators expected co-dictators’ choice by choice.

exp. no. of A choices	recipient CDT	recipient MDT	recipient SDT	pass.dict. CDT
0	37.7	6.5	64.5	33.9
1	62.3	29.0	35.5	66.1
2	0.0	64.5	0.0	0.0

For the question see Figure 5 in Appendix A.1.
Note that in the single and computer treatments there is only a single opponent, hence, there can be no more than one Option A choice.

Table 7: Recipients' and passive dictators' expectations of Option A choices [%].

A.7. Recipients' and passive dictators' expected choices

Recipients, and if present passive dictators were asked for their guess on which option the dictators will choose.[36] Table 7 summarizes the recipients' and passive dictators' expectations. Recipients expected significantly more selfish choices per dictator in MDT (*p*-value[37] 0.0001) and CDT (*p*-value 0.0017) than in SDT but expected fewer selfish choices per dictator in CDT than in MDT (*p*-value 0.0544). The passive dictators' expectations are shown in the right column in Table 7. More than half of the passive dictators expected the dictator in the CDT to choose the selfish option.

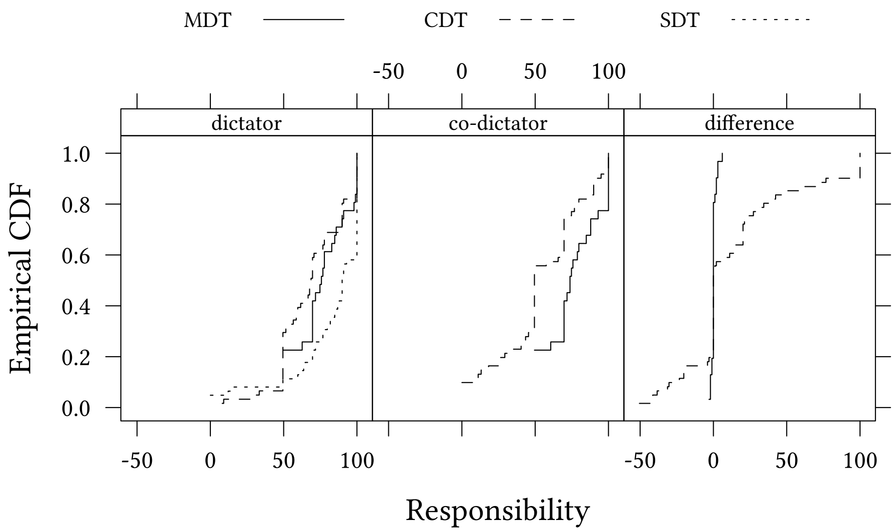
A.8. Recipients' and passive dictators' assigned responsibility to the dictator(s) for the outcome

Recipients, and if present passive dictators were asked how responsible they perceive the human dictator to be for an unfair outcome. Recipients in the MDT and in the CDT were also asked how responsible they perceive the either human or computer co-dictator to be.[38] The allocated responsibility was measured by using a continuous scale from *"Not responsible at all"* (0) to *"Very responsible"* (100). As the left part of Figure 14 shows, recipients assigned a significantly higher level of responsibility to the dictator in the SDT than to the dictators in the CDT (*p*-value 0.0056). However, recipients did not perceive the dictators in the MDT to be significantly less responsible than dictators in the SDT (*p*-value 0.2084).

Perhaps not surprisingly, as the middle part of Figure 14 shows, a human dictator in the MDT was on average perceived as significantly more responsible for the final outcome than the computer in the CDT by recipients (*p*-value 0.0000). Furthermore, as the right part of Figure 14 shows, the allocated responsibility differed more between the human and the computer dictator in the CDT than between the two human dictators in the MDT (*p*-value 0.0034).

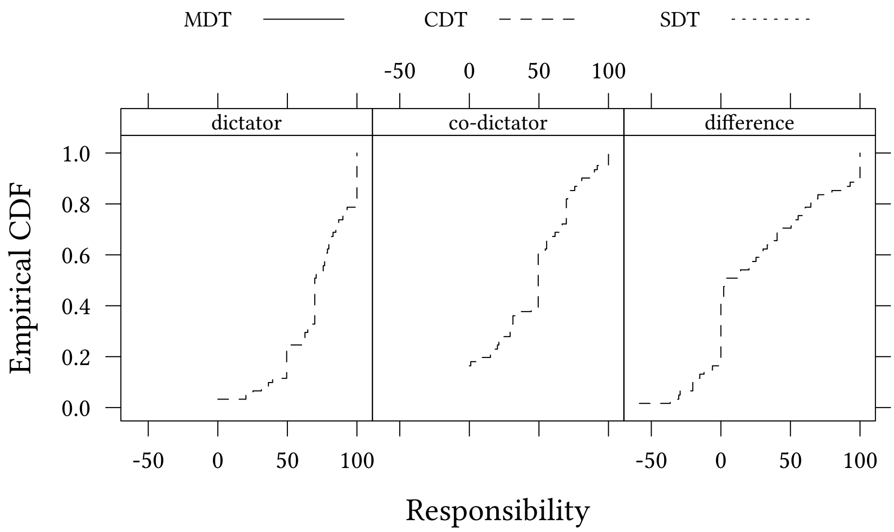
As the left part of Figure 15 shows, a large proportion of the passive dictators perceived the dictator to be very responsible for the final decision. As the middle part of Figure 15 shows, the computer was also perceived as responsible for the outcome. In the right part of Figure 15 we compare the responsibility assigned to the dictator with the responsibility assigned

[36]For the binary Dictator Game interface shown to the recipients see Appendix A.1.1.
[37]The *p*-values in this paragraph are based on a logistic model.
[38]For the exact wording of the question see Question 9 from Appendix A.1.2.



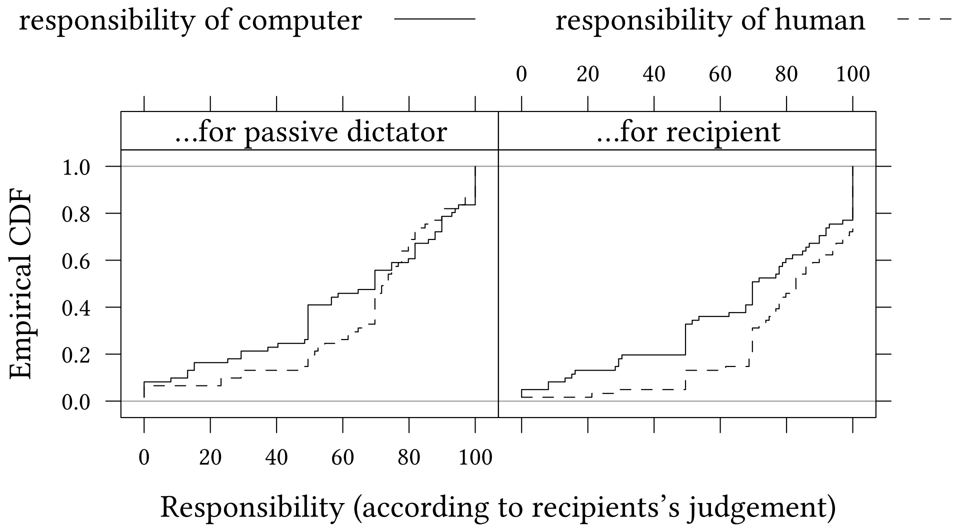
"Dictator" and "co-dictator" are Question 9 from Appendix A.1.2, "difference" shows the difference between the responsibility allocated to the dictators and the co-dictators.

Figure 14: Dictators' responsibility according to recipients.



"Dictator" and "co-dictator" are Question 9 from Appendix A.1.2, "difference" shows the difference between the responsibility allocated to the dictators and the co-dictators.

Figure 15: Dictators' responsibility according to passive dictators.



"Passive dictator" is Question A.1.2 and "recipient" is Question 7 from Appendix 6.

Figure 16: Dictators' personal responsibility and the computer's responsibility for the recipient's and the passive dictator's payoff according to recipients.

to the computer. It becomes clear that a large proportion of the passive dictators hold the dictator far more responsible for the final outcome than the computer (p-value 0.0000).

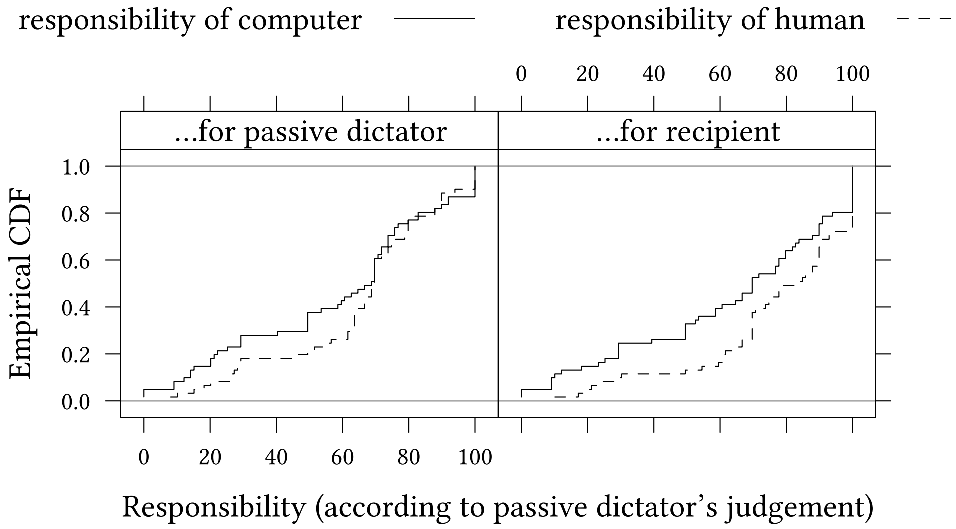
A.9. Recipients' and passive dictators' assigned responsibility to the human dictator(s) for the co-dictators' and the recipients' payoff

Recipient, and if present passive dictators, were asked to evaluate how responsible they perceive the dictator(s) to be for the payoff of the recipient and, if present, the active or passive co-dictator's payoff.[39] The assigned responsibility was measured by using a continuous scale from *"not responsible at all"* (0) to *"totally responsible"* (100).

As Figure 16 shows, recipients in the CDT stated that they perceive the human dictator to be more responsible for the final payoff of the passive dictator as well as for the payoff of the recipient than the computer.

By looking at the difference between the responsibility for the payoff of the passive dictator, see Figure 17, it becomes clear that passive dictators did not perceive the dictator to be significantly more responsible than the computer (p-value 0.1594). However, passive dictators hold the dictator more responsible for the recipient's payoff than the computer (p-value 0.0119).

[39]For the exact wording of the question see Question 6 and Question 7 from Appendix A.1.2.



"Passive dictator" is Question 7 and "recipient" is Question 6 from Appendix A.1.2.

Figure 17: Dictators' personal responsibility and computers' responsibility for the recipient and the passive dictator according to passive dictators.

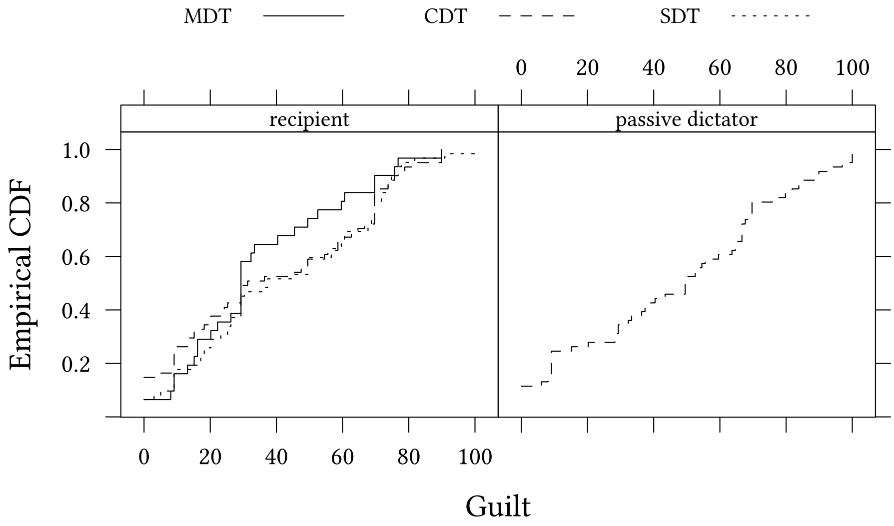
A.10. Recipients' and passive dictators' assigned guilt to the human dictator(s)

In all treatments recipients, and if present passive dictators, were asked to state how guilty they expect the dictators to feel in case Option A would be implemented.[40] The assigned level of guilt was measured by using a continuous scale from *"not guilty"* (0) to *"totally guilty"* (100). Figure 18 shows the anticipated guilt the recipients expected the dictators to perceive in case Option A would be implemented. Recipients in the MDT did not expect the dictators to feel more guilty than recipients in the SDT (p-value 0.2037) or in the CDT (p-value 0.4673) did.

A.11. Manipulation check

A manipulation check was conducted in all treatments. The wording of the manipulation check in the MDT was *"Imagine, now the decision of Player X [Y] is made by a computer. The likelihood the computer chooses Option A (Player X receives 6 ECU, Player Y receives 6 ECU, and Player Z receives 1 ECU) or Option B (Player X receives 5 ECU, Player Y receives 5 ECU, and Player Z receives 5 ECU) is as high as the likelihood experimental subjects chose Option A or Option B in a previous experiment. Example: If three out of ten participants in a previous experiment, whose decision affected the payment, chose a particular option, the computer would choose that option with a probability of 30%. The participants in the previous experiment were not told that their decision would affect a computer's decision in this experiment. Please compare*

[40]For the exact wording of the question see Question 8 from Appendix A.1.2.



"Recipient" and "passive dictator" are Question 8 from Appendix A.1.2.

Figure 18: Dictators' guilt according to recipients and passive dictators.

this decision-making situation with the one Player X and Player Y are confronted with in this experiment.". The wording of the manipulation check in the CDT was *"Imagine, now the decision would not be made by a computer but by Player Y[X] him/herself. Please compare this decision situation to the situation you were confronted with in this experiment."*. The wording of the manipulation check in SDT was *"Imagine, now the decision of Player X is made by a computer."*
As an example, the questions for Player X used in the MDT manipulation check are presented:

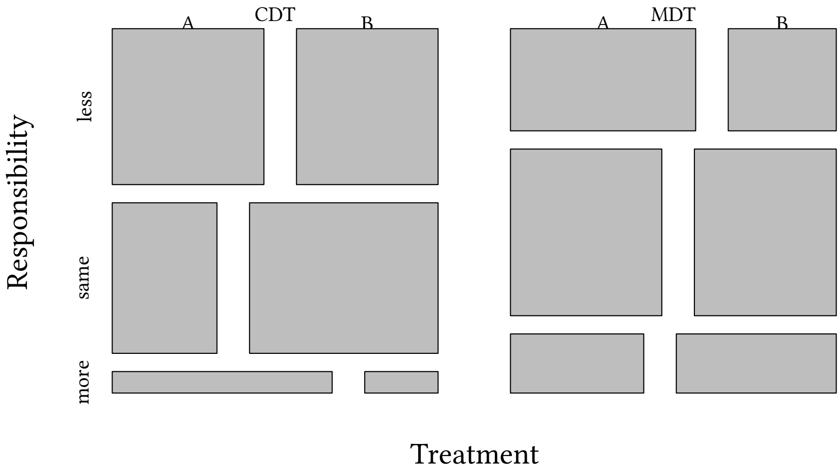
1. How responsible would you feel in this situation for the payoff of Player Y? [Radio buttons *"As responsible as in the experiment"*; *"More responsible than in the experiment"*; *"Less responsible than in the experiment"*] (for an analysis of the answers given see Appendix A.11.1)[41]

2. How responsible would you feel in this situation for the payoff of Player Z? [Radio buttons *"As responsible as in the experiment"*; *"More responsible than in the experiment"*; *"Less responsible than in the experiment"*] (for an analysis of the answers given see Appendix A.11.2)[42]

3. How guilty would you feel if you and the computer both chose Option A and therefore Option A (Player X receives 6 ECU, Player Y receives 6 ECU, Player Z receives 1 ECU) had been implemented? [Radio buttons *"As guilty as in the experiment"*; *"More guilty*

[41]Recipients and passive dictators were asked how responsible they would perceive the dictator to be for the payoff of Player Y in this case.

[42]Recipients and passive dictators were asked how responsible they would perceive the dictator to be for the payoff of Player Z in this case.



For the question see Question 1 from Appendix A.11.

Figure 19: Change in dictators' responsibility for the co-dictator or passive dictator in the manipulation check by dictators.

than in the experiment"; *"Less guilty than in the experiment"*] (for an analysis of the answers given see Appendix A.11.3)[43]

4. Option A will be implemented if you and the computer choose Option A. In this case, Player X receives 6 ECU, Player Y receives 6 ECU and Player Z receives 1 ECU. Please adjust the slide control, so that it shows your perceived responsibility as well as the responsibility you assign to the computer if Option A is implemented. [Slider from *"I am responsible"* to *"I am not responsible"* and slider from *"The computer is fully responsible"* to *"The computer is not responsible"*] (for an analysis of the answers given see Appendix A.11.4 and A.11.5)[44]

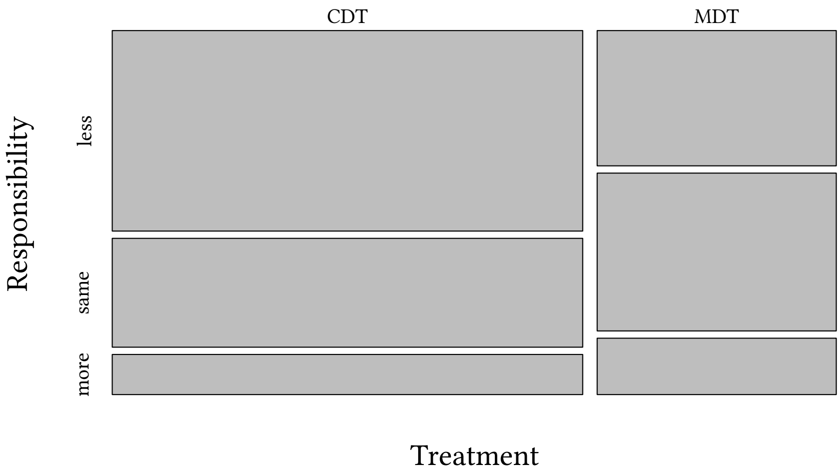
A.11.1. Responsibility for the co-dictator or passive dictator

Results for dictators are shown in Figure 19. Perhaps not surprisingly, dictators in the CDT who imagined sharing their decision with a human instead of a computer stated to feel less responsible for the payoff of their co-dictator (p-value from a binomial test 0.0000). However, dictators in the MDT who imagined sharing their decision with a computer did not feel more responsible for the payoff of the other dictator (p-value from a binomial test 0.2005).

Results for recipients are shown in Figure 20. Recipients in the CDT expected that dictators, who would now have to decide with another human instead of with a computer, to feel

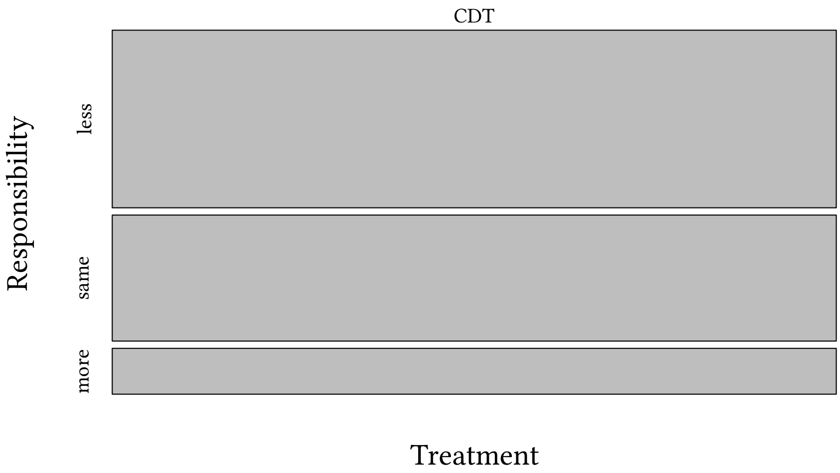
[43]Recipients and passive dictators were asked how guilty they would expect the dictator to feel in case Option A would be implemented.

[44]Recipients and passive dictators were asked how responsible they would expect the dictator to feel in case Option A would be implemented.



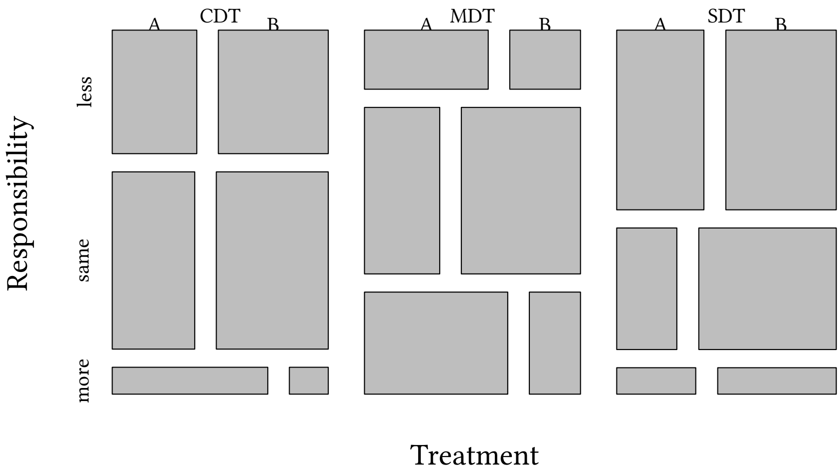
For the question see Question 2 from Appendix A.11.

Figure 20: Change in dictators' responsibility for the co-dictator or passive dictator in the manipulation check by recipients.



For the question see Question 2 from Appendix A.11.

Figure 21: Change in dictators' responsibility for the passive dictator in the manipulation check by passive dictators.



For the question see Question 2 from Appendix A.11.

Figure 22: Change in dictators' responsibility for the recipient in the manipulation check by dictators.

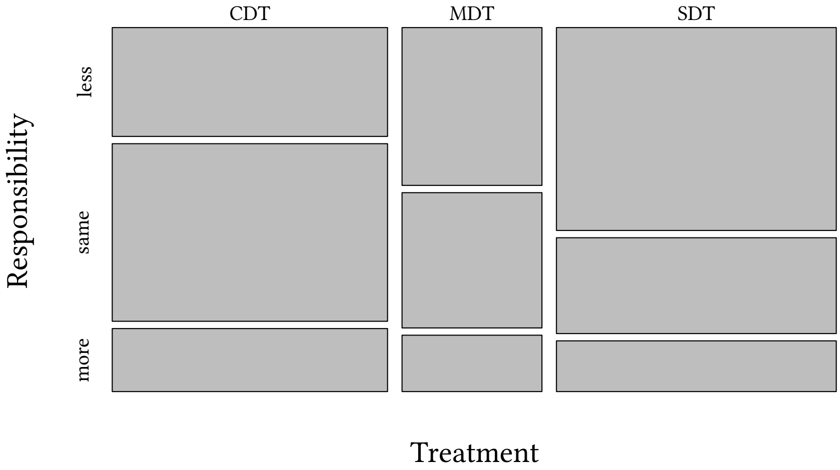
significantly less responsible for the payoff of their co-dictator than in the experiment before (p-value from a binomial test 0.0000). However, recipients in the MDT did not expect the dictators, who would now have to decide with a computer instead of with another human, to feel significantly more responsible for the payoff of their co-dictator than before (p-value 0.1435).

Results for passive dictators are shown in Figure 21. Passive dictators expected the dictators to feel significantly less responsible if they were making their decision with another human dictator instead of with a computer (p-value from a binomial test 0.0003).

A.11.2. Responsibility for the recipient

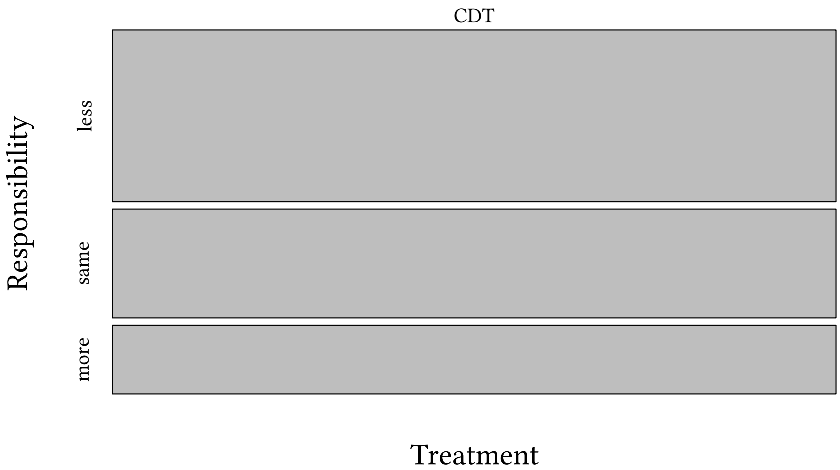
Results for dictators are shown in Figure 22. Dictators in the CDT perceived themselves to be less responsible for the payoff of the recipient once they decide together with a human instead of a computer (p-value from a binomial test 0.0009). Dictators in the MDT did not feel significantly more responsible for the payoff of the recipient once their human counterpart would be replaced with a computer (p-value from a binomial test 0.2005). Dictators in the SDT felt significantly less responsibility for the payoff of the recipient if the decision would be made by a computer and not by themselves in the manipulation check (p-value from a binomial test 0.0000).

Results for recipients are shown in Figure 23. Recipients in the CDT did not expect the dictators, who would have to share their decision with a human instead of a computer, to feel less responsible for the recipients than before (p-value from a binomial test 0.2005). However, recipients in the MDT, expected the dictators, who share their decision with a computer instead of another human, to feel less responsible for the recipients' payoff (p-value from a



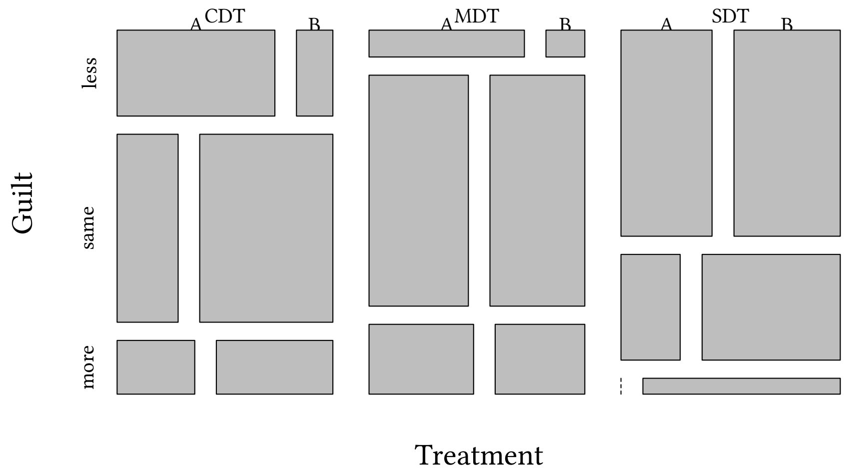
For the question see Question 2 from Appendix A.11.

Figure 23: Change in dictators' responsibility for the recipient in the manipulation check by recipients.



For the question see Question 2 from Appendix A.11.

Figure 24: Change in dictators' responsibility for the recipient in the manipulation check by passive dictators.



For the question see Question 3 from Appendix A.11.

Figure 25: Change in the dictators' perceived guilt in the manipulation check by dictators.

binomial test 0.0636). Recipients in the SDT expected the dictator to feel significantly less responsible for the recipients' payoff if the decision would be made by a computer (p-value from a binomial test 0.0001).

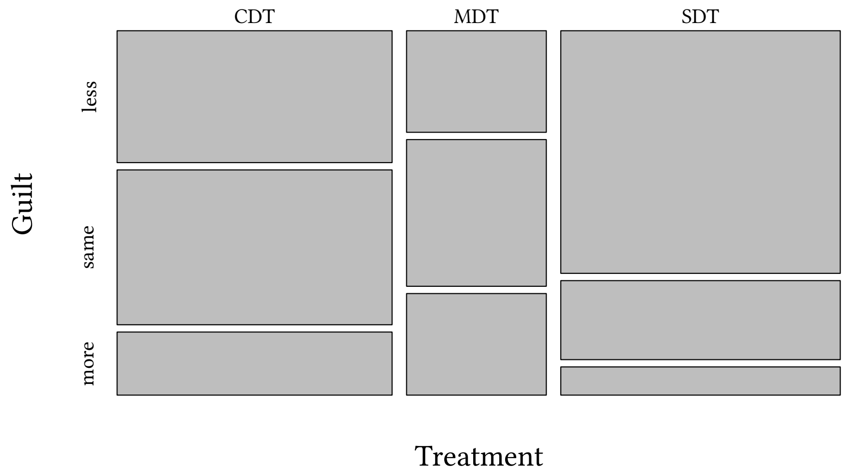
Results for passive dictators are shown in Figure 24. Passive dictators expected the dictator to feel less responsible for the payoff of the responder, if the dictator would decide together with another human instead of with a computer (p-value from a binomial test 0.0079).

A.11.3. Perceived guilt

Results for dictators are shown in Figure 25. Dictators in the CDT stated to feel less guilty once they would be able to share the decision with a human instead of a computer. However, the effect is not significant (p-value from a binomial test 0.3269). Dictators in the MDT did not feel significantly more guilty once their human counterpart was hypothetically replaced with a computer (p-value from a binomial test 0.0963). However, dictators in the SDT stated that they would feel significantly less guilty if the decision would have been made by a computer (p-value from a binomial test 0.0000).

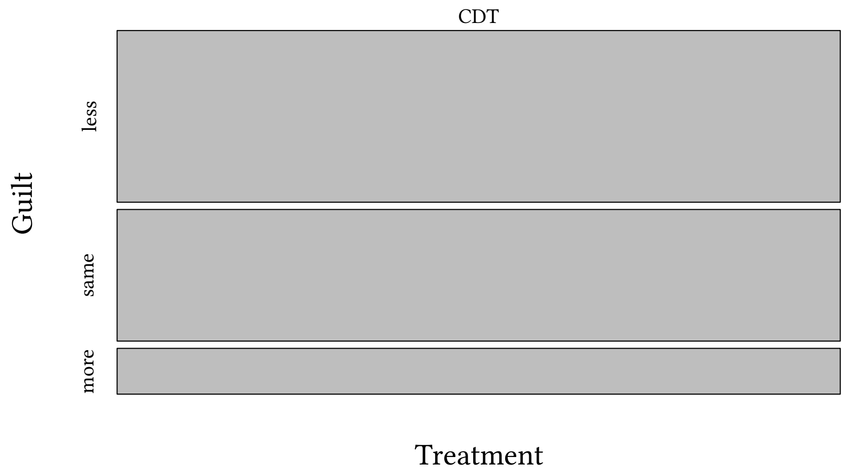
Results for recipients are shown in Figure 26. Recipients in the CDT expected the dictators to feel less guilty when they are sharing the decision with another human (p-value from a binomial test 0.0576). However, in the MDT the number of recipients expected the dictators to feel more guilty or less guilty when deciding together with a computer instead of with another human was nearly evenly distributed (p-value from a binomial test 1.0000). Recipients in the SDT expected the dictators to feel less guilty if the decision would be made by a computer and not by the dictator himself/herself (p-value from a binomial test 0.0000).

Results for passive dictators are shown in Figure 27. Passive dictators expected that the dictators feel less guilty, if they would have to decide together with another human than



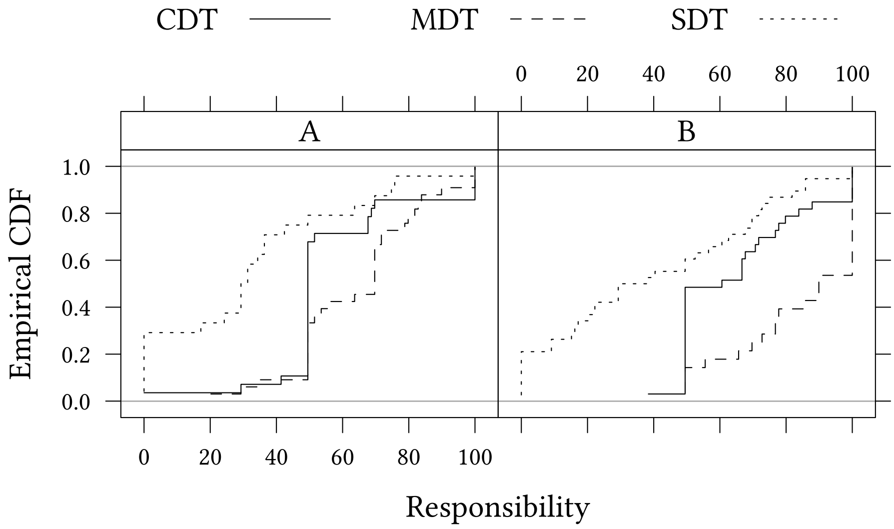
Question 3 from Appendix A.11.

Figure 26: Change in the dictators' perceived guilt in the manipulation check by recipients.



Question 3 from Appendix A.11.

Figure 27: Change in the dictators' perceived guilt in the manipulation check by passive dictators.



For the question see Question 4 from Appendix A.11.

Figure 28: Dictators' personal responsibility in the manipulation check by dictators.

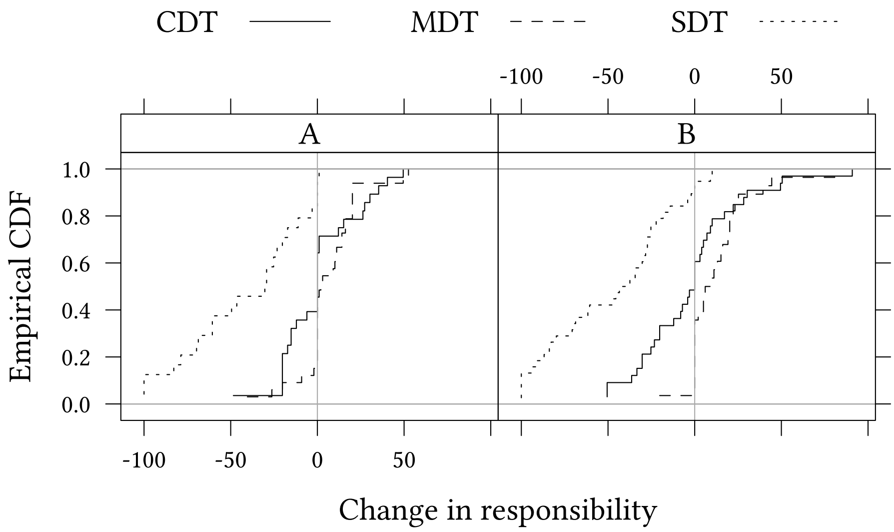
when they decide together with a computer (p-value from a binomial test 0.0005).

A.11.4. Dictators' perceived personal responsibility and assigned responsibility to a human dictator or a computer

The personal responsibility perceived by the dictators in the manipulation check is shown in Figure 28. As could have been expected, dictators in the SDT claimed to perceive themselves to be not very responsible if the decision would have been made by a computer. Interestingly, dictators in the CDT felt less responsible for the final payoff if they had to decide with another human dictator than dictators in the MDT imagining to have to decide with a computer (p-value 0.0022).

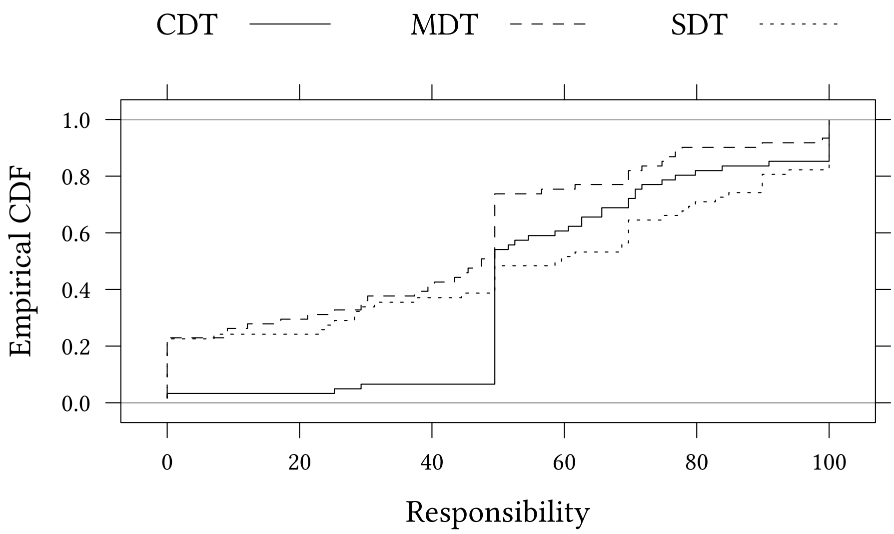
For a comparison of the relative change between the perceived personal responsibility in the hypothetical situation and the perceived personal responsibility in the actual experiment by choice see Figure 29. Dictators in the SDT stated that they would feel less responsible if a computer was to decide on their behalf (p-value 0.0000). Furthermore, the perceived personal responsibility increased for dictators in the MDT when they imagine their counterpart to be replaced by a computer (p-value 0.0260). However, the perceived personal responsibility did not decrease significantly for dictators in the CDT when their counterpart was hypothetical replaced by a human (p-value 0.8388). As the right part of Figure 29 shows, this was mainly driven by dictators who chose Option B.

The responsibility assigned to the co-dictator by the dictators in the manipulation check is shown in Figure 30. While in the SDT the computer's responsibility was assigned equally, significantly more responsibility was assigned to a hypothetical human dictator in the CDT manipulation check than to a hypothetical computer dictator in the MDT manipulation check (p-value 0.0002).



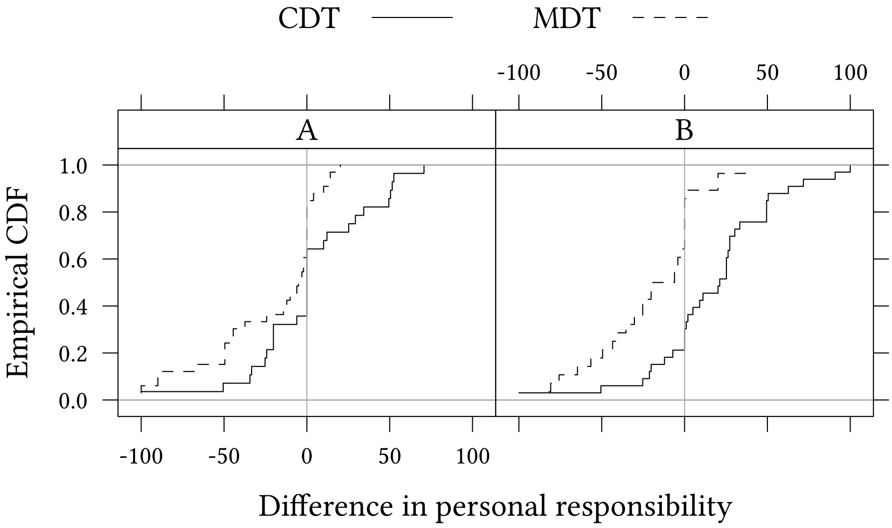
The figure shows the difference between the personal responsibility in the hypothetical situation (described in Appendix A.11) and the actual experiment (as shown in Figure A.1.2).

Figure 29: Dictators' personal responsibility: manipulation check vs. experiment by dictators.



Question 4 from Appendix A.11.

Figure 30: Responsibility assigned to the computer or human co-dictator in the manipulation check by dictators.



The figure shows the difference in the personal responsibility assigned by the dictator to the human or computer co-dictator between the hypothetical situation (described in Appendix A.11) and the actual experiment (as shown in Figure A.1.2).

Figure 31: Responsibility assigned to the computer or human co-dictator: manipulation check vs. experiment by dictators.

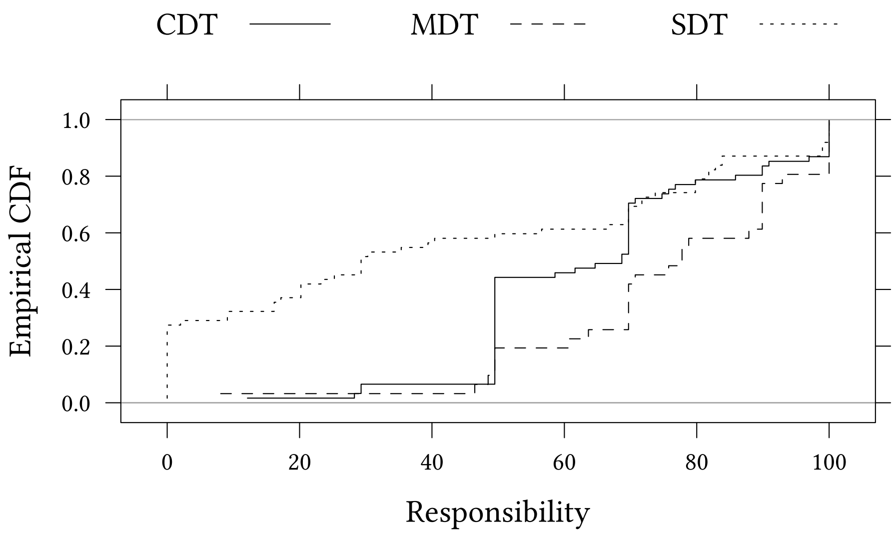
The increase or decrease in the responsibility assigned to the other dictator between the hypothetical situation and the actual experiment by choice is shown in Figure 31. The responsibility attributed to the co-dictator in the CDT increased significantly once the other player is no longer a computer but a human (p-value 0.0392). Similarly, responsibility decreases significantly in the MDT once the other player is no longer a human but a computer (p-value 0.0000). As Figure 31 shows, this was even stronger for dictators who chose Option B.

A.11.5. Recipients' and passive dictators' assigned responsibility to a human dictator or a computer

The responsibility of the dictator(s) for the final payoff perceived by recipients in the manipulation check is shown in Figure 32. Recipients in the SDT perceived the dictators to be not very responsible if the decision had been made by a computer. Furthermore, recipients in the CDT, where the switch was made from a computer to human co-dictator, perceived the dictators to be less responsible for an unfair outcome than the recipients in the MDT, where the switch was made from a human to computer co-dictator, did (p-value 0.0298).

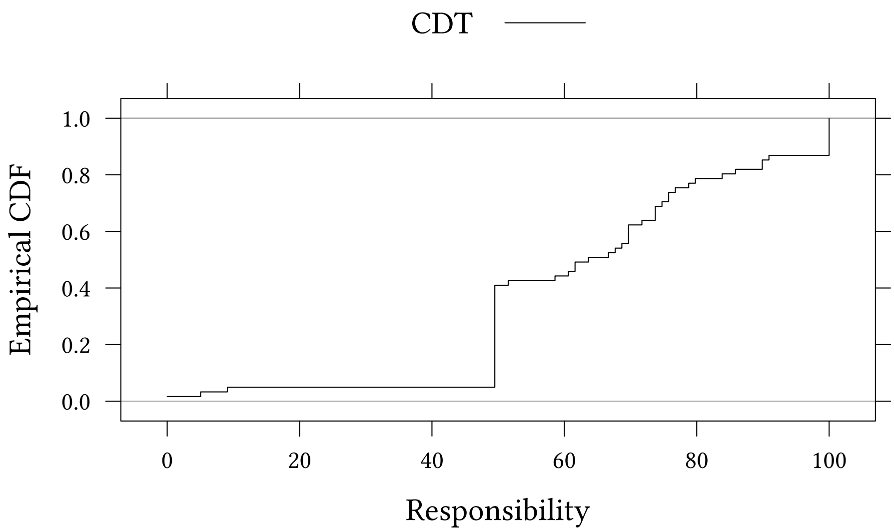
The responsibility of the dictator for the final payoff perceived by passive dictators in the manipulation check is shown in Figure 33. Passive dictators perceived the dictators to be also quite responsible when deciding together with another human.

For a comparison of the relative changes in the recipients' perception of the responsibility of the dictator(s) for an unfair outcome in the hypothetical situation and in the actual exper-



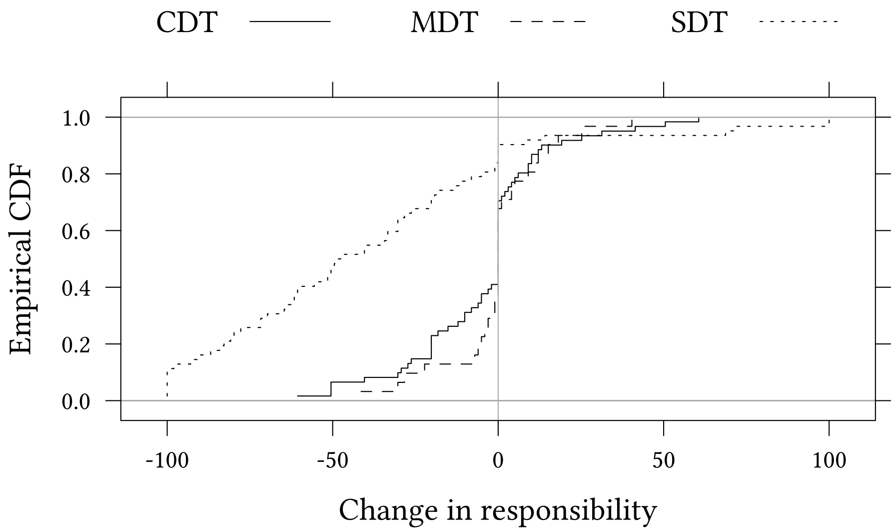
Question 4 from Appendix A.11.

Figure 32: Dictators' personal responsibility in the manipulation check by recipients.



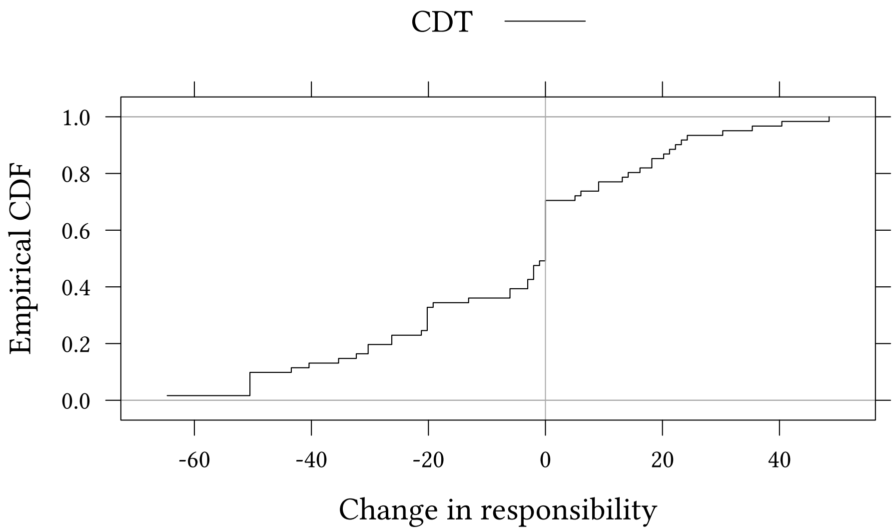
Question 4 from Appendix A.11.

Figure 33: Dictators' personal responsibility in the manipulation check by passive dictators
.



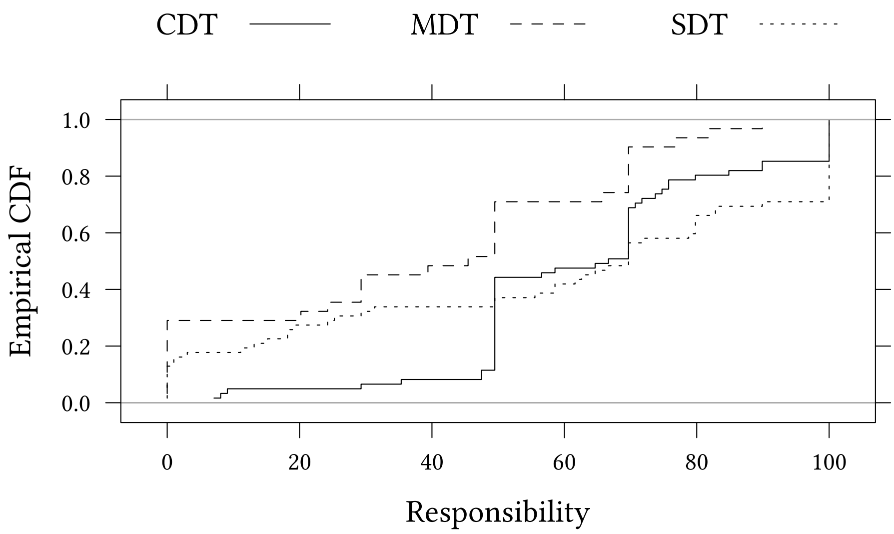
The figure shows the difference in the personal responsibility that the recipients assign to the dictator(s) for an unfair outcome between the hypothetical situation (described in Appendix A.11) and the actual experiment (as shown in Figure A.1.2).

Figure 34: Dictators' personal responsibility: manipulation check vs. experiment by recipients.



The figure shows the difference in the personal responsibility that the passive dictator expect the dictator to perceive for the decision between the hypothetical situation (described in Appendix A.11) and the actual experiment (as shown in Figure A.1.2).

Figure 35: Dictators' personal responsibility: manipulation check vs. experiment by passive dictators.



For the question see Question 4 from Appendix A.11.

Figure 36: Responsibility assigned to the computer or human co-dictator in the manipulation check by recipients.

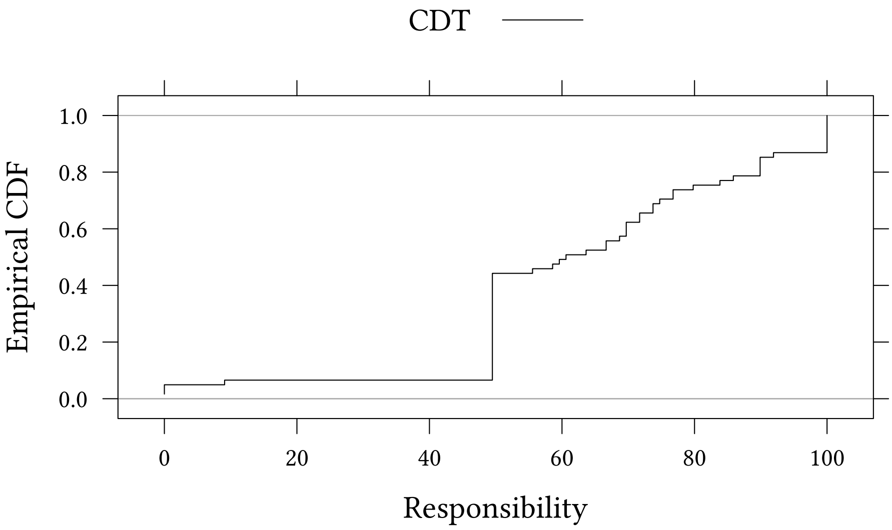
iment see Figure 34. Recipients in the SDT assigned less responsibility for an unfair outcome to the dictator when a computer was to decide on their behalf (p-value 0.0000). However, recipients did not perceive the dictators to be significantly more responsible for an unfair outcome in the MDT when their counterpart was hypothetically replaced by a computer (p-value 0.9590). The same applies for the CDT, where recipients did also not perceive the dictators to feel less responsible for an unfair outcome if the computer would be replaced by a human dictator (p-value 0.3054).

For a comparison of the relative changes between the perceived responsibility of the dictator(s) for the outcome in the hypothetical situation and in the actual experiment by passive dictators see Figure 35. A large but not significant proportion of the passive dictators perceived the dictator to be less responsible if their counterpart is a human instead of a computer (p-value 0.1382).

The responsibility assigned by the recipients in the manipulation check to the either human or computer co-dictator is shown in Figure 36. A computer that decides on its own which option will be implemented, as in the SDT, is perceived as significantly more responsible by the recipients as a computer that determined the final outcome together with a human dictator, as in the MDT, (p-value 0.0031). In addition, the human dictator in the CDT was also perceived as more responsible for an unfair outcome than the computer in the MDT (p-value 0.0001).

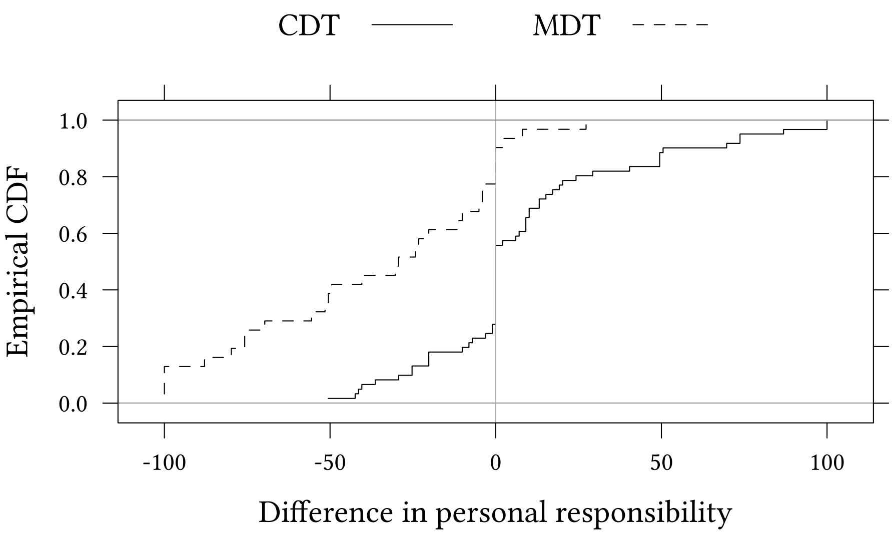
The responsibility assigned by the passive dictators in the manipulation check to the either human or computer co-dictator is shown in Figure 37. Passive dictators perceived both human dictators to be responsible to the same extent for the final outcome (p-value 0.8159).

For a comparison of the relative change in the recipients' perception of the responsibility



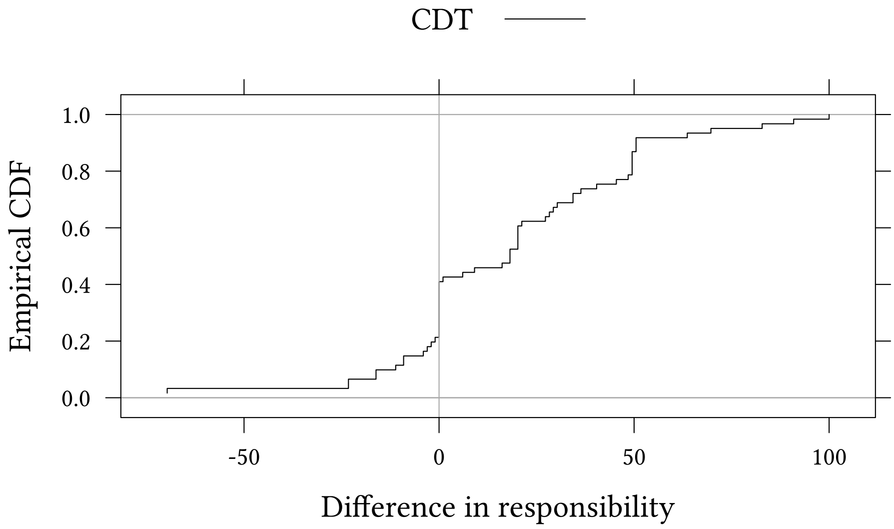
For the question see Question 4 from Appendix A.11.

Figure 37: Responsibility assigned to the human co-dictator in the manipulation check by passive dictators.



The figure shows the difference in the personal responsibility assigned by the recipients to the computer or human dictator for an unfair outcome between the hypothetical situation (described in Appendix refapp:interface:-questions-manip) and the actual experiment (as shown in Figure A.1.2).

Figure 38: Responsibility assigned to the computer or human co-dictator: manipulation check vs. experiment by recipients.



The figure shows the difference in the personal responsibility assigned by the passive dictator to the human dictator between the hypothetical situation (described in Appendix A.11) and the actual experiment (as shown in Figure A.1.2).

Figure 39: Responsibility assigned to the computer or human co-dictator: manipulation check vs. experiment by passive dictators.

of the co-dictator(s) for an unfair outcome in the hypothetical situation and the actual experiment see Figure 38. Recipients in the MDT assigned significantly less responsibility for an unfair outcome to the computer in the manipulation check than they assigned to the human dictator in the actual experiment (p-value 0.0000). Correspondingly, recipients in the CDT assigned significantly more responsibility to the human dictator for an unfair outcome in the manipulation check than they assigned to the computer in the actual experiment (p-value 0.0483).

For a comparison of the relative changes in the passive dictators' responsibility assigned to the human dictator(s) in the hypothetical situation and the computer in the actual experiment see Figure 39. Passive dictators perceived a hypothetical human dictator in the manipulation check to be significantly more responsible for the final outcome than the computer in the actual experiment (p-value 0.0003).

A.12. Effect sizes

All in all we use 399 participants in our study. We find some significant effects, but we also see in many respects differences between the perception of human and computer interaction partners are very small. In this appendix we ask what effect sizes one could expect and how many participants are used in other studies.

Table 8 summarises several studies that also measure responsibility and guilt. Where these studies measure guilt or responsibility not on a scale from 0 to 100 (but on a scale from 1 to 7, 1 to 9, or 1 to 11) we have translated the effect size to a scale from 0 to 100.

Study	Treatment	Dependent variable	n	Effect size on a scale from 0-100
Forsyth et al. (2002)	group size	personal responsibility	122	38.9
Burnette and Forsyth (2008)	sucess × openness	own responsibility	96	7.2
Whyte (1991)	group vs. invidual	personal responsibility	173	22.4
Savitsky et al. (2005)	group vs. invidual	average responsibility	52	8.8
Gosling et al. (2006)	choice	responsibility	51	18.8
Botti and McGill (2006)	choice	responsibility	96	54.2
Mynatt and Sherman (1975)	group × win	responsibility	80	17.8
de Melo et al. (2016)	human/computer	guilt	140	3.5
Reuben and van Winden (2010)	destruction (period 2)	guilt	55	11.7

Table 8: Effect sizes of other studies measuring responsibility and guilt.

Tables 9–11 show the required effect sizes to reach significance for the tests provided in Tables 2–6. The calculation of the required effect sizes for different levels of significance α and power p is based on the (within each group) variance and size of the actual sample.

We see that the studies reported in Table 8 all use fewer participants than we use in our study. We also see that the most of the effect sizes reported in Table 8 should lead to a significant result with the tests reported in Tables 9–11.

	SDT–MDT				CDT–MDT			
		$\alpha =$ 0.01	$\alpha =$ 0.05	$\alpha =$ 0.1		$\alpha =$ 0.01	$\alpha =$ 0.05	$\alpha =$ 0.1
outcome	$p = 0.7$	10.29	8.22	7.17	$p = 0.7$	10.45	8.34	7.27
	$p = 0.8$	11.34	9.27	8.21	$p = 0.8$	11.51	9.41	8.34
	$p = 0.9$	12.80	10.72	9.67	$p = 0.9$	13.00	10.89	9.81
						$\alpha =$ 0.01	$\alpha =$ 0.05	$\alpha =$ 0.1
payoff co-dictator					$p = 0.7$	13.41	10.71	9.34
					$p = 0.8$	14.78	12.08	10.71
					$p = 0.9$	16.69	13.98	12.60
		$\alpha =$ 0.01	$\alpha =$ 0.05	$\alpha =$ 0.1		$\alpha =$ 0.01	$\alpha =$ 0.05	$\alpha =$ 0.1
payoff recipient	$p = 0.7$	10.04	8.02	6.99	$p = 0.7$	11.21	8.95	7.81
	$p = 0.8$	11.07	9.04	8.02	$p = 0.8$	12.36	10.10	8.95
	$p = 0.9$	12.49	10.46	9.43	$p = 0.9$	13.95	11.68	10.53

Table 9: Required effect sizes to reach significance for tests from Table 2.

SDT–MDT				CDT–MDT			
	$\alpha =$ 0.01	$\alpha =$ 0.05	$\alpha =$ 0.1		$\alpha =$ 0.01	$\alpha =$ 0.05	$\alpha =$ 0.1
$p = 0.7$	13.41	10.71	9.34	$p = 0.7$	13.78	11.01	9.60
$p = 0.8$	14.78	12.08	10.70	$p = 0.8$	15.19	12.41	11.00
$p = 0.9$	16.68	13.97	12.60	$p = 0.9$	17.15	14.36	12.95

Table 10: Required effect sizes to reach significance for tests from Table 3.

CDT–MDT			
	$\alpha =$ 0.01	$\alpha =$ 0.05	$\alpha =$ 0.1
$p = 0.7$	-10.51	-7.98	-6.63
$p = 0.8$	-11.69	-9.14	-7.80
$p = 0.9$	-13.31	-10.76	-9.41

Table 11: Required effect sizes to reach significance for tests from Table 6.