# Responsibility, Long-Term Decisions and AI-Interaction: Three Essays on Social Preferences

**Doctoral thesis for obtaining the**

**academic degree Doctor of Economics**

**(Dr. rer. pol.)**

submitted by

Regina Stumpf

at the

Universität Konstanz

Faculty of Politics, Law and Economics

Department of Economics

Konstanz, 2024

Tag der mündlichen Prüfung: 17. Juli 2024

1. Referent: Prof. Dr. Urs Fischbacher
2. Referent: Prof. Dr. Sebastian Fehrler
3. Referent: Prof. Dr. Wolfgang Gaissmaier

# Danksagung

Mit der Abgabe meiner Dissertation neigen sich nicht nur fast fünf Jahre der Promotion, sondern auch elf Jahre an der Universität Konstanz allmählich dem Ende zu. Auf dieser Reise haben mich viele Menschen begleitet, denen ich zutiefst dankbar bin.

Zunächst möchte ich mich bei meinem Betreuer, Urs Fischbacher, bedanken. Seit meinem Bachelor arbeite ich nun an Urs' Lehrstuhl und habe all meine Abschlussarbeiten innerhalb des Lehrstuhls verfasst. Ich habe in dieser Zeit, vor allem in der Promotion, viel von Urs gelernt und wurde immer fachlich und menschlich unterstützt. Urs hat sich immer Zeit für meine Fragen und Anliegen genommen und mir in jeder Hinsicht hilfreiches Feedback in meiner Promotion gegeben. Dabei ist klar, dass Urs ein herausragender Forscher ist. Seine zwischenmenschlichen Qualitäten und seinen freundlichen Umgang möchte ich mit den Worten von Jan Hausfeld zum Ausdruck bringen: "Great researchers are often just described as great researchers. In Urs' case, however, others do not describe him as a great researcher (as everyone knows this is true), but they describe Urs as a nice guy." Urs, vielen Dank für alles!

Des Weiteren möchte ich mich bei Sebastian Fehrler bedanken, der mein Zweitbetreuer ist und mich immer freundlich und offen unterstützt hat. Dabei war es für Sebastian immer selbstverständlich, eine schnelle Lösung für meine Fragen und Anliegen zu finden und niemals den Spaß bei der Arbeit zu verlieren.

Außerdem danke ich meinem Drittbetreuer Wolfgang Gaissmaier. Besonders hervorheben möchte ich ein Seminar, das ich bei Wolfgang belegt habe, wodurch ich viele spannende Forschungsthemen der Sozialpsychologie kennengelernt habe. Die Teilnahme am Seminar hat dann auch zu einer tollen Grillparty für alle Teilnehmer*innen des Seminars und seine Mitarbeiter*innen geführt – danke nochmals dafür!

Ich bin dankbar für all meine weiteren Co-Autoren (Deepti Bhatia, Jan Hausfeld, Baiba Renerte und Fabian Dvorak), ohne die ich meine Dissertation nicht geschafft hätte. Es hat mir großen Spaß gemacht, mit euch an den Forschungsprojekten zusammenzuarbeiten.

Meine Promotionszeit wurde auch maßgeblich durch das Team des Lehrstuhls und des Thurgauer Wirtschaftsinstitutes bereichert. Meine vielen Jahre im Team zeigen schon, wie wohl ich mich gefühlt habe. Ich danke euch für produktives Feedback zu Forschungsideen, Präsentationen und Forschungsarbeiten. Aber die Zeit wäre nur halb so schön gewesen, ohne die vielen lustigen Mittagessen, Retreats, Fondue-Essen, Kaffeepausen und Bürogespräche!

Ein herzliches Dankeschön geht natürlich an all meine Freunde – und insbesondere an unsere eingeschworene Trash-Gruppe! Viele von uns hatten innerhalb ihrer Promotion oder ihrer Jobs Höhen und Tiefen in den letzten Jahren, aber es gab nichts, was nicht ein Nachmittag im Biergarten oder ein tolles Hüttenwochenende nicht wiedergutgemacht hätte. Danke, für all die schönen Erlebnisse mit euch!

Außerdem möchte ich mich bei meiner Familie bedanken. Sie hat mich auf meiner Reise an der Uni immer unterstützt und mir zugesichert, dass ich es schaffe. Der bedingungslose Rückhalt (ohne wirklich zu verstehen, was ich eigentlich genau mache) war mir eine große Hilfe. Und nun habe ich es auch (fast) geschafft!

Mein größter Dank gilt meinem Mann, Flo. Wir sind zusammen vor elf Jahren nach Konstanz gezogen und haben hier unseren ersten gemeinsamen Lebensabschnitt begonnen. Wir haben so viel gemeinsam erlebt, sind hier (ein Stück weit) erwachsen geworden und ich bin von Herzen dankbar, dich an meiner Seite zu haben. Ich freue mich sehr auf den nächsten Lebensabschnitt nach der Promotion, der durch unsere eigene kleine Familie bald beginnt!

Konstanz, Mai 2024

# Contents

# Summary

Social preferences are an inherent characteristic of the majority of people and describe the tendency not only to care about oneself but also to consider others (positively and negatively) in our decision-making (Fehr and Charness, 2023). Our behavior is often driven by our social preferences, which we express in actions such as praising or blaming politicians, donating to a charity, or trusting a friend.

This thesis, entitled "*Responsibility, Long-Term Decisions and AI-Interaction: Three Essays on Social Preferences*", consists of three independent research papers. Each chapter examines how social preferences are influenced by various aspects using an experimental approach. Studying social preferences by means of a controlled laboratory or online experiment has several advantages. First, economic experiments are helpful in mimicking real-world problems of social preferences that might not otherwise be studied due to complexity or a lack of observational data. Second, experiments can be used to study differences in social preferences between different samples and shed light on the heterogeneity of behavior. Finally and most importantly, controlled economic experiments allow causal inferences to be drawn about the influence of various aspects on social preferences.

## Chapter 1: Responsibility

The first research paper addresses the topic of responsibility attribution for collective decisions and poses the research question: "*How do people attribute responsibility when an outcome is not caused by an individual but results from a decision chain involving several people?*". In our society, good and bad outcomes are often the result of the interaction of many people. The production of a firm's good, for example, depends on the sequential efforts of many departments, which ultimately makes it difficult to identify those responsible for success or failure. In a laboratory experiment, we study the motives of responsibility attribution by allowing people to allocate blame and praise in the form of punishment and reward for a sequential group decision made by others. The group decision consists of individuals voting sequentially for a fair or unfair outcome, which is determined by majority rule.

We find that responsibility is linked to the individual choices of the group members, as unfair choices are punished and fair choices are rewarded. Furthermore, people attribute more responsibility to group members who had an impact on the final outcome, and especially to the pivotal group member. People are heterogeneous in how they attribute responsibility, with three

general patterns: Low attribution, Pivotality-driven attribution, and Choice-driven attribution. The results suggest that different motives play a role in how people attribute responsibility for collective decisions.

## Chapter 2: Long-Term Decisions

The second research paper considers different interventions to foster long-term decisions to answer the research question: "*What encourages individuals and groups to envision the long-term future and act on that basis?*". In our society, we are often faced with the problem of balancing and prioritizing current and future issues. How should managers and board members of an organization develop a long-term strategy for the company while focusing on quarterly indicators? How should politicians tackle the climate crisis while dealing with short-term election campaigns? In many cases, a long-term vision is essential to solving long-term challenges.

In two experimental settings – a laboratory experiment with students and an online experiment with a representative sample of the German working population – we let participants decide between investing in real-world projects with either a short time horizon of one year or a long time horizon of up to a decade. We implement different decision-making mindsets to study a shift in people's short-term investments towards long-term investments and focus on mental imagery, the act of creating visual images in the mind's eye.

Our results show that a brief and standard mental imagery intervention leads to more long-term decisions compared to other established mindsets. Our proposed conceptual framework suggests that the creation of detailed images of future outcomes affects the perception of the time horizon of long-term projects, which in turn leads to more long-term decisions. This mechanism works particularly well for people who are optimistic about their future life satisfaction. Furthermore, we show that long-term decisions in groups are mainly determined by the individual preferences of the group members. The findings suggest that mental imagery may be a useful intervention to shift people's current decisions towards more long-term outcomes.

## Chapter 3: AI-Interaction

The third research paper deals with the consequences of generative artificial intelligence (AI) on social preferences and addresses the research question: "*How do people respond to the use of AI in social interactions?*". Not least since the public launch of ChatGPT in November 2022, the use of AI in our daily lives has increased sharply and is now indispensable. This form of AI can efficiently take over many of our decisions and tasks: writing applications, solving customer problems or planning our next vacation. What remains unclear is how people react when AI is used in a social interaction with another human, taking over the decision of the interaction partner. We focus on the role of transparent or opaque use of AI and study the welfare consequences.

We conduct a large online experiment in which human participants interact in various two-person economic games covering different aspects of social preferences. In these interactions, participants either directly interact with another participant or with ChatGPT acting on behalf of the interaction partner. We vary whether participants know about the nature of their interaction partner. We also study the delegation behavior of participants to AI. Furthermore, we conduct a second experiment to investigate whether ChatGPT's decisions can be distinguished from human decisions.

Our results demonstrate that transparent interaction with an AI induces welfare-reducing decisions in humans and has detrimental effects on social preferences, including reduced fairness, trust, trustworthiness, cooperation and coordination. However, these negative effects disappear when people do not know whether their interaction partner is a human or an AI. This finding is puzzling for two reasons: First, people often delegate decisions to AI and believe it is appropriate. Second, people make welfare-reducing decisions when interacting with AI, even though the AI's decisions are, on average, indistinguishable from human decisions. Our results suggest that human skepticism towards AI cannot be solved by increasing transparency alone, as this may paradoxically have the most pronounced negative welfare effects.

# Zusammenfassung

Soziale Präferenzen sind ein inhärentes Merkmal der Mehrheit der Menschen und beschreiben die Tendenz, sich nicht nur um sich selbst zu kümmern, sondern auch andere (positiv und negativ) in unsere Entscheidungsfindung einzubeziehen (Fehr and Charness, 2023). Unser Verhalten wird häufig durch unsere sozialen Präferenzen bestimmt, die wir in Handlungen wie Lob oder Tadel für Politiker, Spenden für eine Wohltätigkeitsorganisation oder Vertrauen in einen Freund zum Ausdruck bringen.

Diese Arbeit mit dem Titel "*Verantwortung, langfristige Entscheidungen und KI-Interaktion: Drei Aufsätze über soziale Präferenzen*" besteht aus drei unabhängigen Forschungsarbeiten. In jedem Kapitel wird anhand eines experimentellen Ansatzes untersucht, wie soziale Präferenzen durch verschiedene Aspekte beeinflusst werden. Die Untersuchung sozialer Präferenzen mit Hilfe eines kontrollierten Labor- oder Online-Experiments hat mehrere Vorteile. Erstens sind ökonomische Experimente hilfreich, wenn es darum geht, reale Probleme sozialer Präferenzen nachzuahmen, die andernfalls aufgrund ihrer Komplexität oder eines Mangels an Beobachtungsdaten nicht untersucht werden könnten. Zweitens können Experimente genutzt werden, um Unterschiede in den sozialen Präferenzen zwischen verschiedenen Stichproben zu untersuchen und die Heterogenität des Verhaltens zu erhellen. Und schließlich, was am wichtigsten ist, lassen kontrollierte ökonomische Experimente kausale Rückschlüsse auf den Einfluss verschiedener Aspekte auf soziale Präferenzen zu.

## Kapitel 1: Verantwortung

Das erste Forschungspapier befasst sich mit dem Thema der Verantwortungszuweisung bei kollektiven Entscheidungen und stellt die Forschungsfrage: "Wie weisen Menschen die Verantwortung zu, wenn ein Ergebnis nicht von einer Einzelperson verursacht wird, sondern das Resultat einer Entscheidungskette ist, an der mehrere Personen beteiligt sind?". In unserer Gesellschaft sind gute und schlechte Ergebnisse oft das Resultat einer Interaktion vieler Menschen. Die Produktion eines Unternehmensgutes zum Beispiel hängt von den aufeinanderfolgenden Bemühungen vieler Abteilungen ab, was es letztlich schwierig macht, die Verantwortlichen für Erfolg oder Misserfolg zu ermitteln. In einem Laborexperiment untersuchen wir die Motive der Verantwortungszuweisung, indem wir Menschen die Möglichkeit geben, Schuld und Lob in Form von Bestrafung und Belohnung für eine sequentielle Gruppenentscheidung, die von anderen getroffen wurde, zuzuweisen. Die Gruppenentscheidung besteht darin, dass die Indi-

viduen nacheinander für ein faires oder unfaires Ergebnis abstimmen, das schlussendlich durch die Mehrheitsregel bestimmt wird.

Wir stellen fest, dass die Verantwortung mit den individuellen Entscheidungen der Gruppenmitglieder verbunden ist, da unfaire Entscheidungen bestraft und faire Entscheidungen belohnt werden. Außerdem schreiben die Menschen denjenigen Gruppenmitgliedern mehr Verantwortung zu, die einen Einfluss auf das Endergebnis hatten, und insbesondere dem zentralen Gruppenmitglied. Menschen sind heterogen in der Art und Weise, wie sie Verantwortung zuweisen, wobei es drei allgemeine Muster gibt: geringfügige, pivotal- und entscheidungsorientierte Verantwortungszuschreibung. Die Ergebnisse deuten darauf hin, dass unterschiedliche Motive eine Rolle dabei spielen, wie Menschen die Verantwortung für kollektive Entscheidungen zuweisen.

# Kapitel 2: Langfristige Entscheidungen

Das zweite Forschungspapier befasst sich mit verschiedenen Interventionen zur Förderung langfristiger Entscheidungen, um folgende Forschungsfrage zu beantworten: "Was ermutigt Individuen und Gruppen, sich die langfristige Zukunft vorzustellen und auf dieser Grundlage zu handeln?". In unserer Gesellschaft sind wir oft mit dem Problem konfrontiert, zwischen aktuellen und zukünftigen Problemen abzuwägen und Prioritäten zu setzen. Wie sollen Führungskräfte und Vorstandsmitglieder einer Organisation eine langfristige Strategie für das Unternehmen entwickeln, während sie sich gleichzeitig auf vierteljährliche Indikatoren konzentrieren? Wie sollen Politiker*innen die Klimakrise angehen, während sie sich mit kurzfristigen Wahlkämpfen beschäftigen? In vielen Fällen ist eine langfristige Vision unerlässlich, um langfristige Herausforderungen zu lösen. In zwei experimentellen Studien - einem Laborexperiment mit Studierenden und einem Online-Experiment mit einer repräsentativen Stichprobe der deutschen Erwerbsbevölkerung - lassen wir die Teilnehmenden zwischen Investitionen in reale Projekte mit entweder einem kurzen Zeithorizont von einem Jahr oder einem langen Zeithorizont von bis zu einem Jahrzehnt entscheiden. Wir setzen verschiedene Entscheidungsmentalitäten ein, um eine Verschiebung der kurzfristigen Investitionen der Menschen hin zu langfristigen Investitionen zu untersuchen. Wir konzentrieren uns auf mentale Bilder, d. h. auf die Erzeugung detaillierter Bilder vor dem geistigen Auge.

Unsere Ergebnisse zeigen, dass eine kurze und standardmäßige Intervention mit mentalen Bildern im Vergleich zu anderen etablierten Denkansätzen zu mehr langfristigen Entscheidungen führt. Der von uns vorgeschlagene konzeptionelle Rahmen legt nahe, dass die Schaffung detaillierter Bilder von zukünftigen Ergebnissen die Wahrnehmung des Zeithorizonts langfristiger Projekte beeinflusst, was wiederum zu mehr langfristigen Entscheidungen führt. Dieser Mechanismus funktioniert besonders gut bei Menschen, die optimistisch sind, was ihre zukünftige Lebenszufriedenheit angeht. Außerdem zeigen wir, dass langfristige Entscheidungen in Gruppen durch die individuellen Präferenzen der Gruppenmitglieder bestimmt werden. Unsere Ergebnisse deuten darauf hin, dass mentale Bilder eine nützliche Intervention sein können, um aktuelle Entscheidungen der Menschen in Richtung langfristiger Ergebnisse zu lenken.

# Kapitel 3: KI-Interaktion

Das dritte Forschungspapier befasst sich mit den Auswirkungen der generativen künstlichen Intelligenz (KI) auf soziale Präferenzen und geht der Forschungsfrage nach: "Wie reagieren Menschen auf den Einsatz von KI in sozialen Interaktionen?". Nicht zuletzt seit dem öffentlichen Start von ChatGPT im November 2022 hat der Einsatz von KI in unserem täglichen Leben stark zugenommen und ist nicht mehr wegzudenken. Diese Form der KI kann viele unserer Entscheidungen und Aufgaben effizient übernehmen: das Schreiben von Bewerbungen, das Lösen von Kundenproblemen oder die Planung unseres nächsten Urlaubs. Unklar ist, wie Menschen darauf reagieren, wenn KI in einer sozialen Interaktion mit einem anderen Menschen eingesetzt wird, in der die KI die Entscheidung des Interaktionspartners übernimmt. Wir konzentrieren uns auf die Rolle des transparenten oder intransparenten Einsatzes von KI und untersuchen, welche Konsequenzen dies für die Wohlfahrt hat.

Wir führen ein großes Online-Experiment durch, in dem menschliche Teilnehmende in verschiedenen Zwei-Personen-Wirtschaftsspielen interagieren, die unterschiedliche Aspekte sozialer Präferenzen abdecken. Die Teilnehmenden interagieren entweder direkt mit einer anderen teilnehmenden Person oder mit ChatGPT, das im Namen des Gegenübers handelt. Wir variieren, ob die Teilnehmenden über die Natur des Interaktionspartners Bescheid wissen und untersuchen darüber hinaus auch das Delegationsverhalten der teilnehmenden Personen an die KI. Außerdem, führen wir ein zweites Experiment durch, um zu untersuchen, ob die Entscheidungen von ChatGPT von menschlichen Entscheidungen unterschieden werden können.

Unsere Ergebnisse zeigen, dass eine transparente Interaktion mit einer KI zu wohlfahrtsmindernden Entscheidungen bei Menschen führt und sich negativ auf die sozialen Präferenzen auswirkt, u. a. in Form von weniger Fairness, Vertrauen, Vertrauenswürdigkeit, Kooperation und Koordination. Diese negativen Auswirkungen verschwinden jedoch, wenn die Menschen nicht wissen, ob ihr Gegenüber ein Mensch oder eine KI ist. Dieses Ergebnis ist aus zwei Gründen rätselhaft: Erstens delegieren Menschen häufig Entscheidungen an KI und glauben, dass dies angemessen ist. Zweitens treffen Menschen wohlfahrtsmindernde Entscheidungen, wenn sie mit KI interagieren, obwohl die Entscheidungen der KI im Durchschnitt nicht von menschlichen Entscheidungen zu unterscheiden sind. Unsere Ergebnisse deuten darauf hin, dass die Skepsis der Menschen gegenüber KI nicht allein durch die Erhöhung der Transparenz gelöst werden kann, da dies paradoxerweise die stärksten negativen Auswirkungen auf die Wohlfahrt haben kann.

# Chapter 1

# Blame and Praise: Responsibility Attribution Patterns in Decision Chains

Deepti Bhatia[1,2], Urs Fischbacher[1,2,3], Jan Hausfeld[4,5] and Regina Stumpf[1,2]

[1] Department of Economics, University of Konstanz (Germany)
[2] Thurgau Institute of Economics (Switzerland)
[3] CESifo, Munich (Germany)
[4] CREED and Amsterdam School of Economics, University of Amsterdam (Netherlands)
[5] Tinbergen Institute (Netherlands)

**Abstract**

How do people attribute responsibility when an outcome is not caused by an individual, but results from a decision chain involving several people? We study this question in an experiment, in which five voters sequentially decide on how to distribute money between them and five recipients. The recipients can reward or punish each voter, which we use as measures of responsibility attribution. In the aggregate, we find that responsibility is attributed mostly according to the voters' choices and the pivotality of the decision, but not for being the initial voter. On the individual level, we find substantial heterogeneity with three overall patterns: Little to no responsibility attribution, pivotality-driven, and focus on choices. These patterns are similar when praising voters for good outcomes and blaming voters for bad outcomes.

**Keywords:** Responsibility Attribution, Collective Decision-Making, Voting, Decision Process

**JEL Classification:** C91, C92, D63, D70, D91

**Note:** This is an adapted version of the research article published in *Experimental Economics*. (Bhatia, D., Fischbacher, U., Hausfeld, J., and Stumpf, R. (2024). Blame and praise: Responsibility attribution patterns in decision chains. *Experimental Economics*, 27(3):637–663)

## 1.1 Introduction

How is responsibility attributed when an outcome results from a chain of actions? First, consider a disaster (i.e., sinking of large ships) as an example of a bad outcome. Disasters are often a result of a chain of unfortunate circumstances, decisions, and actions. Whittingham (2004) presents several examples of disasters and discusses the responsibility of the different people involved. Typically, someone makes a mistake which is not detected or appropriately fixed. This mistake causes/adds to further problems until a disaster is unavoidable. Similarly, good outcomes are often the result of the (sequential) interaction of people. For instance, joint production is an example of a positive decision chain. If a firm releases a product, research and development, production, and marketing sequentially contribute to the success. In our study, we experimentally investigate how people attribute responsibility in decision chains by allowing people to allocate blame and praise to others in the form of punishment and reward.[1]

The general question of responsibility attribution has been addressed from different angles. A normative point of view has been taken from a philosophical (Feinberg, 1970) as well as from a legal perspective (Hart and Gardner, 2008). More recently, the question has also attracted the interest of psychologists (Ross and Nisbett, 1991; Weiner, 1995; Gerstenberg et al., 2011) political scientists (Iyengar, 1994) and economists, both from an empirical (Charness, 2000; Bartling and Fischbacher, 2012; Bartling et al., 2015; Duch et al., 2015) and theoretical perspective (Besley, 2006; Bartling and Fischbacher, 2012; Engl, 2022). Our study has an empirical point of view, as we investigate in a lab experiment how people assign responsibility using punishment and reward. Understanding the empirical patterns of responsibility attribution is important because it has consequences for how we set up liability rules and how we distribute the benefits of joint ventures. Our experiment shows what rules of responsibility attribution people spontaneously apply, and the analysis of the heterogeneity provides insights on how well people agree with respect to these rules.

In real life decision chains, the decision makers and the actions differ in many dimensions. In our experiment, we study the impact of the sequence in isolation. For this reason, we investigate decisions in a sequential voting game, in which symmetric voters decide with majority over a good or bad outcome for other people. The subjects in our experiment are matched into groups of five voters and five recipients. The five voters choose sequentially between two options of how to allocate points between voters and recipients, while the outcome is determined by simple majority rule. The two allocations differ in their fairness: the unfair allocation favors

[1] Generally, responsibility is associated with blame and praiseworthiness. According to the Oxford Dictionary (`https://www.oxfordlearnersdictionaries.com/definition/english/responsibility`, last retrieved: 06.03.2024), responsibility comes with three meanings: 1. a duty to deal with or take care of somebody/something, so that you may be blamed if something goes wrong. 2. responsibility (for something); blame for something bad that has happened. 3. a moral duty to do something or to help or take care of somebody because of your job, position, etc. The second meaning and the second part of the first meaning relate responsibility attribution to blame attribution. For this reason, punishment and reward is often used in experiments in order to assess responsibility attribution (Coffman, 2011; Bartling and Fischbacher, 2012; Duch et al., 2015; Gurdal et al., 2013; Oexl and Grossman, 2013; Bartling et al., 2015). There are also other ways to assess responsibility attribution, for example, Engl (2022) directly asks participants about the responsibility of different actors in different scenarios using the Krupka and Weber (2013) method.

the voters, while the fair allocation results in similar payoffs for the voters and recipients. The recipients receive full information about the voting sequence and then have the possibility to sanction the voters. Our treatments differ in the sanctioning options. There is a treatment with only punishment, one with only reward, and one with both. Finally, we use process measures and record the response times for all participants and use eye-tracking for the recipients. We make several contributions to the existing literature. We study different outcomes (good and bad outcomes), compare different sanctioning options (punishment and reward), explore more motives (outcome, choice, intention, initiation, pivotality, causal responsibility based on models), and add process measures (response time and eye-tracking). We are investigating the following research questions:

First, we study *how responsibility is attributed to different roles in the decision chain.* We investigate how the position in the decision chain affects responsibility attribution, and we use two theoretical measures of responsibility as predictors (Bartling and Fischbacher, 2012; Engl, 2022). Two voters are in a particular focus, the first voter in favor of the resulting outcome, which we call initiator and the third voter in favor of the resulting outcome, which we call the pivotal voter.[2] Bartling et al. (2015) show that in sequential decisions, pivotal voters are blamed the most for unfair outcomes. Duch et al. (2015) find that in simultaneous collective decisions, proposal power plays an important role for responsibility attribution. In our design with five voters, we can compare the initiator and the pivotal player with a majority voter who is neither. Further, we can distinguish between majority voters who still had a say and those who had no more influence on the outcome. We call the former the intentional voters and the latter the non-intentional voters.[3] On the aggregate, our experiment shows that punishment targets people who vote for the unfair outcome. People who potentially had an impact on the outcome, and in this respect are intentionally unkind, are punished more, and the pivotal players even more within this group. Analogously, choosing the kind option leads to rewards, which are higher if the choice can be considered as intentional and even higher if the choice is pivotal. We do not find that higher reward or punishment is assigned to the initiator of a good or bad outcome. The theoretical measures of responsibility (Bartling and Fischbacher, 2012; Engl, 2022) are correlated with the empirical responsibility attribution, but the correlations are not very high and these measures do not outperform simpler measures.

Second, we study *whether responsibility attribution differs for good and bad outcomes in comparable situations.* We do so, by making use of our treatments with reward and punishment options, as well as the combination, in which both reward and punishment are available. As mentioned above, the evaluation of responsibility is consistent between reward and punishment. Subjects reward others for good outcomes very similarly to how they punish others for bad

[2]The term pivotality is used in a variety of contexts. In voting decisions, pivotality is also used with respect to the order that is induced by the strength of the preferences in favor of a policy. We use the term with respect to the temporal order. This is meaningful because after the third decision in favor of an option, the outcome of the vote is determined. Our definition also corresponds to the definition in Bartling et al. (2015). A definition of gradual pivotality is used by Engl (2022) as a measure of causal responsibility. We discuss this definition in Section 1.3.

[3]If the decision is already taken, we cannot infer any intention of the voter. For example, if the outcome has already been decided, a vote for the fair outcome does not mean that the player wanted to be kind.

outcomes – both on the aggregate level and in the individual analysis introduced below. The environment in which both reward and punishment are available shows that people tend to prefer to use punishment over reward. However, responsibility attribution is less differentiated compared to the environments in which only one option is available.

Third, as the explanatory power of the general model is not very high, we explore *individual patterns in responsibility attribution*. Studying different patterns in social interactions has been the object of several studies.[4] We find that the individual behavior can be classified into three main groups. These groups are analogous in the reward and the punishment condition. There is a group of subjects who barely rewards or punishes. Another group of subjects particularly targets the pivotal voter, and a third group of subjects mostly attributes responsibility according to the choices of the voters.

Fourth, we investigate *how the voters respond* to the incentives created by the option of punishment and reward. We find evidence that voters are (at least partially) aware of how responsibility is attributed. In particular, they are aware that pivotality matters and partially use delegation in order to avoid blame for unkind decisions or seek responsibility for kind decisions in order to gain credit.

Fifth, we study *the underlying decision process of responsibility attribution.* We analyze the response time patterns of voters and find that they have longer response times when they are potentially pivotal, i.e., if their decision can finalize the outcome. Further, we use eye tracking during the sanctioning decisions. Our results show that the gaze analysis does not confirm the behavioral focus on the pivotal player. If any player is more in the focus, then it is the initiator.

The remainder of this paper is structured as follows: Section 2 explains the experimental design used in this study, while Section 3 outlines the different motives used to study responsibility attribution and lists our predictions. Our results are presented in Section 4 and Section 5 concludes.

## 1.2 Experimental Design

We build on Bartling et al. (2015) who investigate a sequential voting task with punishment. We add a treatment with reward in order to directly compare responsibility attribution for good and bad outcomes, and we increase the number of voters as it allows investigating more roles in this sequential decision process. In our experiment, we randomly assign the role of voter and recipient to subjects. Five voters and five recipients form a group and keep their roles throughout the experiment. The five voters sequentially decide between two allocations in order to distribute 50 points among all ten group members. There are two sets of allocations.

[4]For example, Falk et al. (2008) investigate different patterns of reward and punishment among their subjects. Most participants express both positive and negative reciprocity, while others only show positive or negative reciprocal fairness preferences. Similarly, Leibbrandt and López Pérez (2011) study heterogeneity in costly reward and punishment. Their results indicate that most subjects follow a mixture of outcome-based and reciprocal preferences. Besides observing different patterns of how subjects sanction, Albrecht et al. (2018) go one step further and examine if and how different behavioral patterns are linked within each subject. In a linear public goods game with decentralized punishment, they show that for most subjects cooperation and punishment patterns are aligned.

In one set, voters can choose between a fair allocation, in which all group members receive 5 points, and an unfair allocation, in which the voters receive 9 points each and the recipients receive 1 point each. In the second set, the voters can choose between a fairer allocation, in which the voters receive 6 points and the recipients receive 4 points, and an unfair allocation, in which the voters receive 8 points each and the recipients receive 2 points each. We chose the two sets in such a way that the alternatives create a similar trade-off, and, thus, the two sets can be treated equally.

The position of the voters in the voting sequence is randomly determined. Each voter is informed about the decisions of all previous voters in the sequential process before choosing an allocation. A majority rule is applied, which means that the allocation that is chosen by at least three of the five voters is implemented.

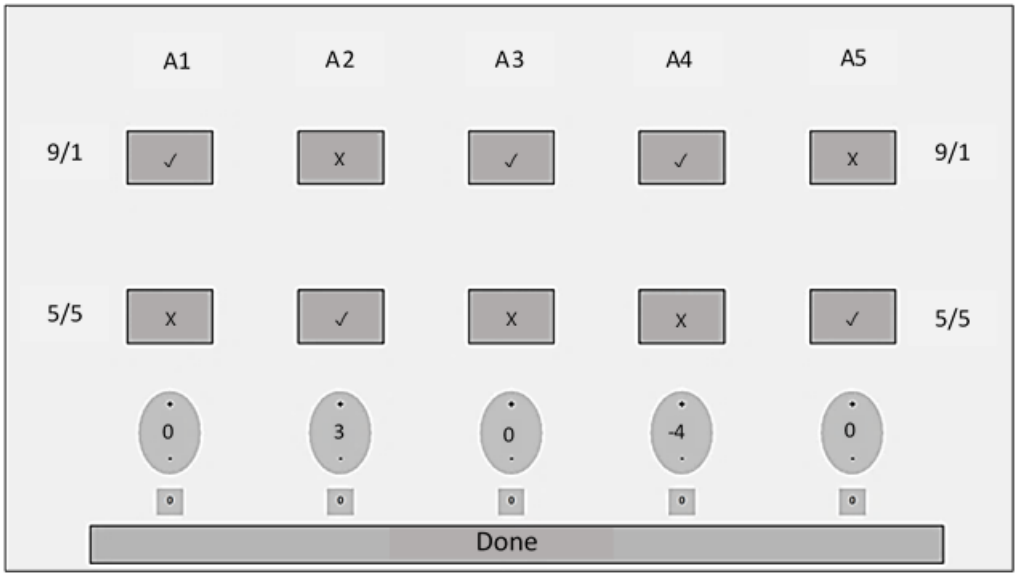


Figure 1.1: Exemplary decision screen of a recipient in the treatment *Both*.

Note: Voters are denoted as A1 - A5 and their decisions are indicated by a check at the selected allocation. The positioning of the allocations in the top or bottom row was randomly determined. The outcome in this example is the unfair allocation (9 points for each voter and 1 point for each recipient) and the recipient attributes three reward points to voter 2 and deducts four punishment points from voter 4. We added the respective allocation on both sides of the screen to minimize subject's gaze being biased towards one side of the screen. The font size in the figure was enlarged for better readability.

The recipients are informed about the individual voting decisions and thus also about the voting outcome. One randomly determined recipient receives an extra point and has the option to sanction the voters. We vary the sanction option across three treatments: Recipients can only punish voters (*Punishment*), they can only reward voters (*Reward*), and they can reward and punish voters (*Both*). In all the treatments, the recipient first has to decide whether to sanction the voters at the cost of the extra point by clicking a button. In the second decision, the recipient can then assign 0 to 7 reward and/or punishment points to each voter individually. Figure 1.1 illustrates an exemplary decision screen of a selected recipient who decided to sanction the voters. The payoff of each voter is determined by the resulting voting outcome and the reward or punishment points the voter receives from the recipient. Each recipient gets a payoff

according to the chosen allocation. The selected recipient can additionally keep the extra point if she decides not to sanction.

The game is played as a one-shot game, and we use the strategy method for both voters and recipients, that is, each voter and each recipient makes choices for all possible scenarios. Each voter chooses between the fair and unfair allocation in every voter position for every possible combination of previous voter choices. This results in 31 binary choices for each of the two allocation sets which we display in random order. Additionally, the voters play one round of a dictator game for each allocation set resulting in two additional decisions. All recipients act as if they were chosen to be the recipient to sanction. For every possible voting sequence, the recipients decide whether they want to sanction any of the voters and if yes, by how much they want to sanction each voter. The scenarios differ in the decision constellations of the voters and the allocation sets.

As process measures, we collect response time data for both voters and recipients. In addition, we use eye-tracking to record the gaze pattern of the recipients to evaluate the information recipients use when attributing responsibility to the voters.

### 1.2.1 Procedural Details

The experiment was programmed using the software "z-Tree" (Fischbacher, 2007). Participants were students who were recruited by the data-system ORSEE (Greiner, 2015). In total, nine sessions were conducted in February 2019, three sessions for each treatment. The experiment was carried out at the experimental laboratory of the University of Konstanz (Lakelab) in Germany. Each session consisted of two groups, 10 voters and 10 recipients, such that there were 30 voters and 30 recipients in each of the three treatments (*Punishment*, *Reward*, *Both*). One subject was excluded from the analysis due to insufficient attention during the session.

The average age of our subjects was 22 years (min: 18, max: 33) and 55.6% of the subjects were female. The subjects earned on average 22.58 EUR (about 25.40 USD at that time) which included a show-up fee and an extra compensation for the usage of eye-tracking. The sessions lasted 90 minutes on average.

We used Tobii EyeX eye-trackers with a sampling rate of 60Hz to record gaze data. The subjects used chin rests to improve data quality, and the seating distance to the screen was approximately 58 cm. The screens were 22-inch color monitors with a resolution of 1920x1080 pixels. The calibration of the subjects to the eye-tracking system was done at the beginning of the experiment via a seven-point calibration. Two additional subjects have been excluded from the eye-tracking analysis because of technical issues and poor quality of the gaze data.

Fixations are identified with the help of the DBSCAN-algorithm (Ester et al., 1996). We create ten non-overlapping areas of interest (AOI), each with a radius of 90 pixels. Each AOI covers a box on the decision screen indicating whether the voter voted in favor or against the specific allocation (see Figure 1.1). Therefore, for each voter there are two AOIs. The horizontal distance between the centers of two AOIs was 320 pixels and the font size of the cues was set to 20.

## 1.3 Criteria and Theoretical Predictions

In this section, we present criteria according to which people could attribute responsibility. It includes motives like intention-based reciprocity and measures of responsibility. We will assess the relevance of these criteria by using them as predictors for reward and punishment. The measures that we will present first have not been intended to be used as measures of responsibility. However, these measures have been suggested to explain reward and punishment. Even when they do not capture all facets of responsibility, they capture some facets, which has also been discussed in the theoretical literature on responsibility. For example, Shaver (1985) mentions dimensions of responsibility, among them, intentions.

**Outcome.** Outcome-oriented models such as Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) are quite popular. In our voting game, it means that reward and punishment is not directed to a specific voter. The assessment only depends on the outcome. The unfair alternative is considered as bad, and the fair alternative as kind or neutral.[5] The models predict equal punishment for all voters in case of an unfair outcome and if at all, equal reward in case of a fair outcome.

**Choice.** This motive assumes that a vote for the fair allocation is perceived as kind and a vote for the unfair allocation as unkind, independent of whether the vote was relevant for the outcome or not. It can be considered as a naive notion of intention. If the voter would not take the behavior of the other voters into account and would believe that their own vote is decisive, then their own vote would correspond to their preference and the vote would express their intention.

**Intention.** This motive captures preferences as suggested in the reciprocity models of Rabin (1993) and Dufwenberg and Kirchsteiger (2004). Since there are only two options, either both are neutral, or one option is kind and the other is unkind. Theoretically, whether voting for fair or unfair is kinder depends on the belief of what the other voters do.[6] However, in the experiment, voting for the fair allocation always results in a higher probability of getting the fair outcome than voting for the unfair allocation. Thus, as long as a majority for one of the allocations is not reached, votes are impactful and voting for the fair allocation can be considered as kind and voting for the unfair allocation as unkind. As soon as the decision is made and the outcome can no longer be changed, votes are no longer impactful, no intentions can be inferred, and the vote is considered as neutral.

**Initiation.** This motive is motivated from Duch et al. (2015) who showed that proposal power is an important aspect in how subjects attribute responsibility for collective decisions.[7] Applied to our experiment, it assumes that the first voter who votes for the resulting outcome has a special responsibility for the outcome.

[5] According to outcome-oriented models, voting for the fair alternative can be considered as neutral and therefore fair outcomes might not be rewarded.

[6] For example, if voters 2 to 5 always vote against voter 1, then the fair vote of voter 1 would actually be unkind.

[7] However, proposal power is not explicit in our experiment. This may limit how people assess the initiator's responsibility.

**Pivotality.** This motive is motivated from Bartling et al. (2015) who found that the pivotal voter is punished more than the non-pivotal voters for voting for the unfair outcome. In our setup, the pivotal voter is the third voter who votes in favor of one of the two allocations. After this choice, the outcome is determined and can no longer be changed.

**Bartling and Fischbacher (2012) Responsibility Measure.** In the responsibility measure formalized by Bartling and Fischbacher (2012) (from now on called BF Responsibility), the responsibility of the different voters for a certain outcome is assigned proportional to how much their vote contributes to an increase in the probability that this outcome results. The measure depends on the belief about the voter's decisions. In order to keep the following explanations simpler, we present the case where the outcome is unfair. The measure works exactly in the same way for the fair outcome. It is calculated as follows: First, we calculate for every decision node the probability that the unfair outcome results. Next, each action (e.g. each vote) gets a raw responsibility, which is the difference between the probability before and after the action. Finally, in order to get the responsibility measure, the raw responsibility is normalized. This means specifically: The responsibility of an action that does not increase the probability of an unfair outcome is set to zero. The responsibility of an action that increases the probability of an unfair outcome is the raw responsibility divided by the sum of all positive raw responsibilities along the decision path from the start to the final outcome. This measure lies between zero and one. As mentioned above, it depends on a belief about the voters' decisions. Practically, we use the empirical distribution of voters' decisions as their belief. Note that the responsibility measure refers to the outcome. We expect that responsibility for the unfair outcome triggers punishment, and responsibility for the fair outcome triggers reward.

**Engl (2022) Responsibility Measures.** Another notion of responsibility has been suggested by Engl (2022). We consider a simple variant of the model and explain what it predicts in our case. Engl (2022) distinguishes between ex-ante and ex-post causal responsibility. For the calculation of ex-post causal responsibility, the outcome is considered as given. The idea of this measure is that the attributed responsibility of an action increases in proportion to the pivotality of this action. An action is considered as pivotal if it causes the outcome, and not choosing it would result in the counter-factual outcome. With respect to the ex-post responsibility, the action has the outcome as consequence and, therefore, the responsibility of the action is just the probability that not choosing this action causes the other outcome.[8] Note that the ex-post responsibility is defined for an action within a decision path. The ex-ante responsibility of a vote is calculated as the expected value of the ex-post responsibility of all paths following this vote. Thus, the ex-ante responsibility fixes the target vote as well as the preceding votes and calculates the expected value of the ex-post responsibilities over all paths that follow the target vote.

Let us illustrate the idea of ex-post responsibility in a few examples where F refers to a vote for the fair allocation, U refers to a vote for the unfair allocation, and the order of F and U refers to the sequence of votes. So, FUFUU represents a situation, in which the first and the

[8]This probability also depends on what the other voters subsequently do. We use the empirical distribution of the voter behavior for this purpose.

third voters vote for fair and the other voters vote for unfair. The fourth and the fifth vote in this sequence are fully responsible for the unfair outcome because choosing fair would result in the fair outcome. In the sequence UFFFF, the fair outcome results. The last voter in this sequence is not responsible for the fair outcome at all because a change of the action does not change the outcome. The ex-post responsibility of the fourth voter in the sequence UFFFF depends on the probability $p$ that the last voter chooses U. If the fourth voter would vote U, then the unfair outcome would result with a probability of $p$. Thus, the ex-post responsibility of voter four for the fair outcome equals $p$. Note that different from the BF measure, the Engl responsibility for the fair outcome is not always zero when the unfair outcome results and vice versa. For example, the first unfair vote (U) in UFFFF does not have zero responsibility for the fair result, and the first voter who votes for F bears some positive responsibility for the unfair outcome. Therefore, we consider the difference between the responsibility for the unfair and the fair outcome as predictors. We use the variables *Ex-ante Engl Difference (U-F)* and *Ex-post Engl Difference (U-F)* as predictors for the *Punishment* treatment and *Ex-ante Engl Difference (F-U)* and *Ex-post Engl Difference (F-U)* as predictors for the *Reward* treatment. Table 1.1 shows a summary of the presented criteria and their theoretical predictions, which we use to analyze how people attribute responsibility. The third column of the table illustrates the theoretical predictions of how responsibility is assigned for the unfair outcome in the voting sequence FUFUU. The fourth column of the table provides the theoretical predictions of the responsibility assignment for the fair outcome in the voting sequence UFFFF. In Section 1.6.3 in the Appendix, we display the responsibility predictions for each voter in each voting sequence according to the models by Bartling and Fischbacher (2012) and Engl (2022).

In the exemplary situation of the voting sequence FUFUU, the unfair outcome results. Therefore, the outcome-based models would predict equal punishment for all voters. If Choice is used as a criterion to assign responsibility, unfair choices will be punished equally (second, fourth and fifth voter) and fair choices will be rewarded equally (first and third voter). The intention-based models would predict equal responsibility for the second, fourth and fifth voter for the unfair outcome since a majority of votes was not reached before. With respect to initiation, the second voter would be solely responsible for the unfair outcome, since this voter was the first to vote for the unfair allocation. The fifth voter in the sequence is the pivotal voter and would be fully responsible according to this criterion. BF Responsibility would predict zero responsibility for the voters who voted for the fair allocation (first and third position voters) and a responsibility between 0 and 1 for the second, fourth and fifth position voter. As outlined above, the ex-post Engl Responsibility model predicts that the fourth and fifth voters are fully responsible. All the other voters are partially responsible. The ex-ante Engl Responsibility is non-zero for all voters because the unfair outcome is possible in all decisions. It equals 1 for the last voter, because for the last voter, the ex-ante and the ex-post measures coincide. For voter 4, the ex-ante measure is the expected ex-post measure of the two sequences FUFUF and FUFUU. For voter 3, it is the expected ex-post measure of four sequences, and so on.

Table 1.1: Overview of Responsibility Measures

| Criterion | Theoretical prediction - Who is responsible? | Responsibility for unfair outcome in | | | | | Responsibility for fair outcome in | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | (F | U | F | U | U) | (U | F | F | F | F) |
| Outcome | All voters are responsible for the final outcome. | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Choice | All voters who vote for the final outcome are responsible. | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| Intention | Impactful voters are more responsible than non-impactful voters. | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0 |
| Initiation | The first voter voting for the final outcome is responsible. | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| Pivotality | The third voter voting for the final outcome is responsible. | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| BF Responsibility | Voters are more responsible if they increase the probability that the final outcome results. | 0 | .29 | 0 | .34 | .37 | 0 | .26 | .51 | .23 | 0 |
| Ex-post Engl Responsibility | Voters are more responsible if choosing the alternative action would have resulted in a different outcome with a higher probability. | .18 | .91 | .17 | 1 | 1 | .35 | .93 | .81 | .40 | 0 |
| Ex-ante Engl Responsibility | Expectation of ex-post Engl Responsibility. Fixes the votes preceding the target vote and takes the expectation with respect to subsequent votes. | .05 | .57 | .05 | .63 | 1 | .06 | .37 | .66 | .40 | 0 |

Summary of the presented criteria and their theoretical predictions. The third and fourth columns shows the predictions of the criteria for each voter in two voting sequences of fair (F) and unfair (U) choices. The third column reports the responsibility for the unfair outcome in FUFUU. The fourth column reports the responsibility for the fair outcome in UFFFF. BF and the Engl measures depend on the distribution of the voting decisions. We show the measures from the punishment treatment in the allocation set $(9, 1)$ vs. $(5, 5)$.

## 1.4 Results

Our main research focus is how recipients use punishment and reward (attribute responsibility) for sequential collective decisions, which are either fair or unfair. To do so, we first test the criteria and theoretical predictions stated in Section 1.3, followed by an integrative model, and a short analysis of individual heterogeneity and process data. In the analysis of the voters' behavior, we study voting patterns, strategic voting and delegation, and response times. We do not distinguish between the two possible allocation settings because the results are, as expected, quite similar.[9][10]

[9] In the dictator games, 21 out of 90 subjects chose (5,5) when (9,1) was the alternative and 20 chose (6,4) when (8,2) was the alternative.

[10] The predictions of BF and Engl depend on the voter behavior. Thus, these predictions depend on the treatments. The analysis uses these treatment specific predictions.

### 1.4.1 Sanctioning Behavior

For the analysis, we separate the decisions according to whether the outcome was fair or unfair, and classify each voter into the majority group (those who voted for the resulting outcome) or into the minority group (those who voted against the resulting outcome). In each voting sequence, there are between three and five majority voters and between zero and two minority voters. Among the majority voters, we distinguish between voters who vote before a majority is reached and who are therefore impactful for the final outcome (first three majority voters) and non-impactful voters (the possible fourth and fifth majority voter). Finally, we separate the impactful voters (first three majority voters) into the initiator, i.e., the first majority voter, the second majority voter and the pivotal voter, who is the third majority voter. The impactful voters are named according to their roles in the sequential decision.

**Outcome.** Figure 1.2 shows the average punishment and reward points for fair and unfair outcomes across treatments. Note that in the *Both* treatment, recipients could use the seven points for both punishment and reward. Thus, the *Both* treatment is included in both subfigures.

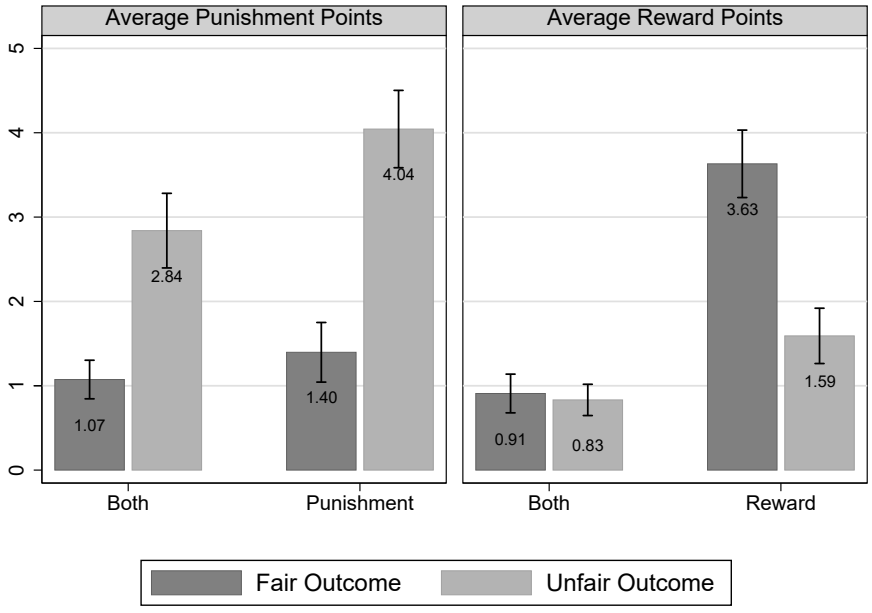


Figure 1.2: Average punishment and reward points for fair and unfair outcomes across treatments.

Note: Standard error bars are shown in black.

In the *Punishment* treatment, recipients punish more in unfair outcomes than in fair outcomes (4.04 and 1.40 points, Wilcoxon signed-rank test, $p < 0.001$).[11] In the *Reward* treatment, more reward points are assigned on average in fair outcomes compared to unfair outcomes (3.63

[11] For every hypothesis test we use a Wilcoxon signed-rank test for matched samples which is based on average decisions per subject.

and 1.59 points, $p < 0.001$). In the *Both* treatment, where subjects can both punish and reward voters, more punishment points are used on average for unfair outcomes (2.84) than for fair outcomes (1.07, $p < 0.001$). In contrast, recipients do not reward fair and unfair outcomes differently (0.91 and 0.83 points, $p = 0.975$). This shows that punishment is used more frequently and in a more differentiated way than reward when both options are available.

Our results show that recipients indeed punish unfair outcomes more than fair outcomes. However, our outcome-based prediction cannot be supported insofar as recipients also punish when the outcome is a fair allocation and use reward points for both outcomes.

**Choices.** Figure 1.3 shows the average sanction points for different voter roles for unfair and fair outcomes across treatments. Table A1.2 in the Appendix lists the corresponding average sanction points in more detail by also taking the voter position into account. It is important to note that not every voter role is equally represented across all possible scenarios due to the natural composition of all possible voting choice constellations. Table A1.3 shows the average sanction points for each voter position in each scenario faced by the recipients.

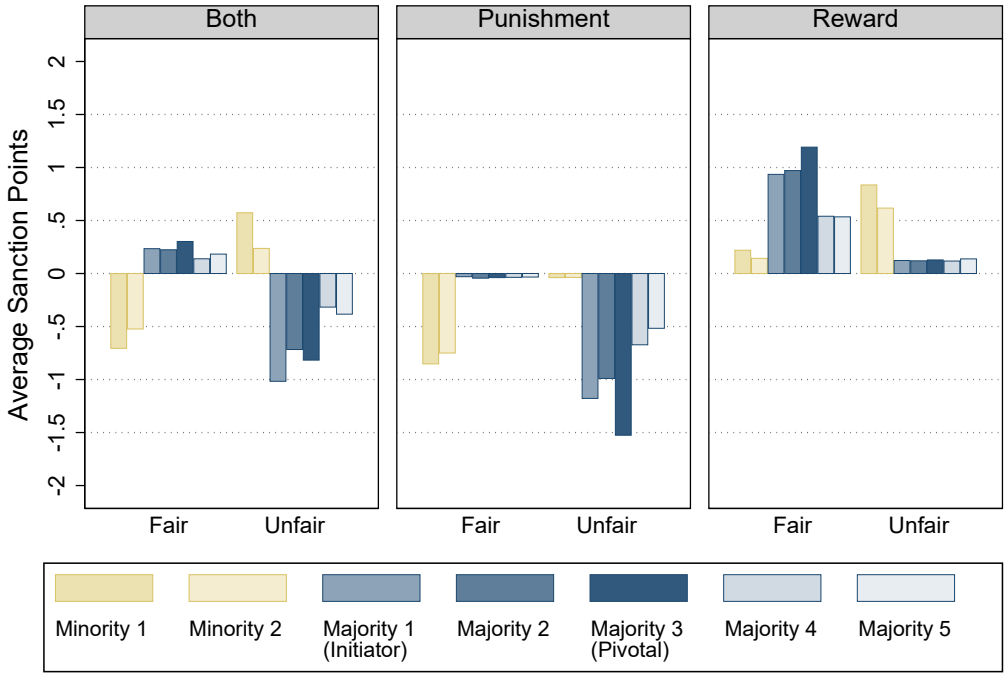


Figure 1.3: Average sanction points for different voter roles across treatments.

Note: The bars show the average sanction points for different sanction motives separated by outcome (fair vs. unfair) and treatment. Depending on the voting sequence the following voter roles are possible: Minority 1 represents the first voter voting against the final outcome. Minority 2 represents the second minority voter. Majority 1 (Initiator) is the first voter to vote for the final outcome. Majority 2 is the second majority voter. Majority 3 (Pivotal) is the third voter to vote for the final outcome. Majority 4 and Majority 5 are the fourth and fifth majority voter.

Figure 1.3 shows that choice clearly matters for responsibility attribution. In all treatments, voting for the fair allocation is rewarded more and/or punished less than voting for the unfair allocation (all *p-values* $< 0.02$). This is illustrated by comparing the minority voters (yellow

bars) with the respective majority voters (blue bars) in Figure 1.3 for each outcome. Our choice-based prediction can thus be confirmed by the data and shows that subjects attribute responsibility according to choices. In particular, there is virtually no punishment for fair votes and, vice versa, hardly any reward for unfair votes. Nevertheless, reward points for subjects voting for the unfair allocation as well as punishment points for subjects voting for the fair allocation are significantly different from zero for almost all voter roles (all *p-values* $< 0.05$, except for one case).

**Intentions.** We now disentangle who is held more responsible among the voters choosing the same allocation. We call the first three majority voters of a voting sequence intentional voters, while the majority voters four and five are non-impactful voters, as their vote can no longer change the outcome. As shown in Figure 1.3, recipients punish intentional voters for unfair outcomes and reward them for fair outcomes more than non-impactful voters (all *p-values* $< 0.05$). In the *Both* treatment, the same results hold except for one case.[12]
Our results show that the intention-based prediction can be confirmed. Thus, when attributing responsibility for an outcome, recipients take the impact of the votes on the final outcome into account.

**Initiation.** We analyze whether there are differences in responsibility attribution among intentional voters. We first test whether the initiator is sanctioned more than the second majority voter. Across all treatments, we do not find evidence for an initiator effect (average sanction points for initiator vs. second majority voter: *Punishment* -1.18 vs. -0.99; *Reward* 0.94 vs. 0.97; *Both* fair outcomes 0.23 vs. 0.22; *Both* unfair outcomes -1.02 vs. -0.72). Recipients do not seem to punish and/or reward the initiator differently than the second majority voter (all *p-values* $> 0.1$). Therefore, the initiation-based prediction does not hold.

**Pivotality.** We expect recipients to attribute the highest responsibility to the pivotal voter (e.g., Bartling et al. (2015)). In the *Punishment* treatment, the pivotal voter is punished more than both other intentional voters when the outcome is unfair (-1.52 vs. -1.18 / -0.99, both *p-values* $< 0.02$). In the *Reward* treatment, the pivotal voter is rewarded the most for fair outcomes (1.19 points on average) which is more than the other two intentional voters (both *p-values* $< 0.08$). However, in the *Both* treatment, the pivotal voter is not treated differently compared to other intentional voters (all *p-values* $> 0.1$). The pivotality-based prediction can partially be confirmed. Pivotality plays an important role when attributing responsibility for cases where reward and punishment are separately available.
Taken together, our analysis allows us to answer the first two research questions on how responsibility is attributed to different roles in a decision chain and between different outcomes. Our results show that people attribute responsibility differently depending on the outcome of the sequential decision. Generally, subjects are held responsible according to their choices. Furthermore, impactful voters are perceived to be more responsible than non-impactful voters. Last, the pivotal voter bears the highest responsibility, while the initiator of a sequential voting sequence is not treated differently than the second impactful voter. In addition, we find that

[12]For fair outcomes, the second majority voter and the fifth majority voter are not treated differently.

the different criteria of responsibility attribution are the same whether people praise others for good outcomes or blame them for bad outcomes.

### Econometric Comparison of Sanctioning Motives

In this section, we provide an econometric comparison between the different motives and measures of responsibility attribution. As the correlation tables in Section 1.6.4 of the Appendix show, there are high correlations between the different motives and measures. Therefore, in order to compare the explanatory power of the different motives, we first study them in isolation.

Figure 1.4 shows the $R^2$ of individual OLS regressions with reward and/or punishment points as dependent variables and the criteria of responsibility attribution as presented in Section 1.3 as independent variables.[13]

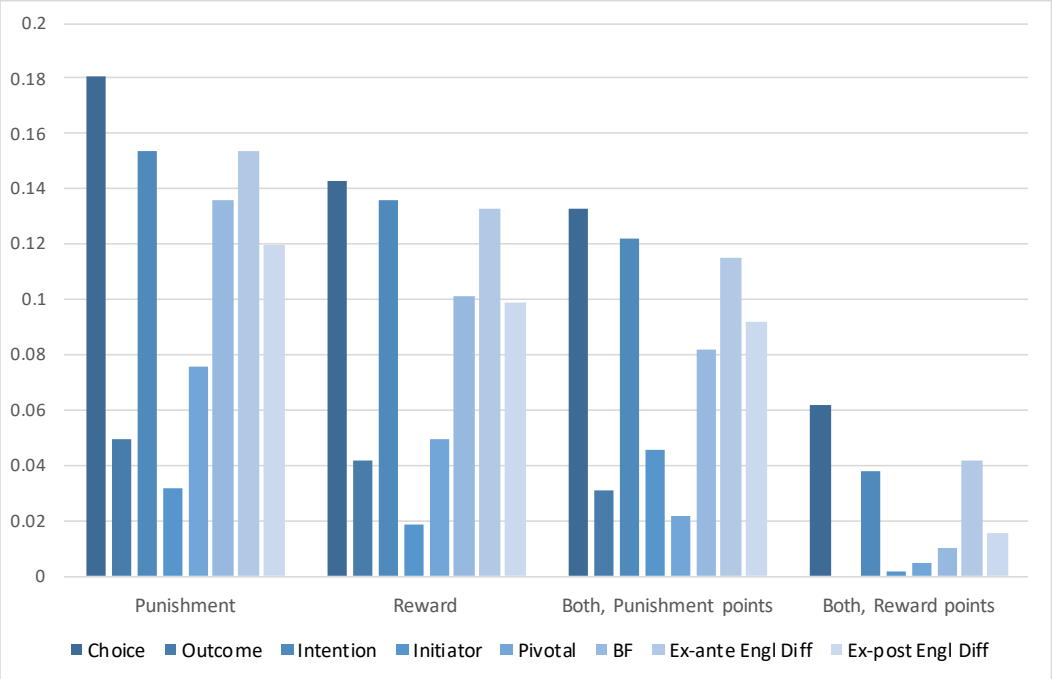


Figure 1.4: Comparison of $R^2$ for different responsibility measures

The figure shows that the criterion Choice has the highest explanatory power across all treatments, followed by the criterion Intention, further followed by the theories of responsibility attribution (BF, Ex-ante Engl Diff, Ex-post Engl Diff), which have similar explanatory power but among them the Ex-ante Engl Diff measure has the highest explanatory power in all situations. The criteria Pivotality and Initiation are amongst the criteria with the least explanatory power.[14]

So far, each motive of how responsibility is attributed has been tested separately. Outcome-based models predict when people use punishment or reward, but they do not predict who

[13]The output of the individual OLS regressions can be found in Section 1.6.5 of the Appendix. Note, that when taken individually, each responsibility measure significantly predicts the attributed sanction points.

[14]With respect to explanatory power, the theories of responsibility attribution might be seen as a good compromise between the individual motives. Overall, the explanatory power is not very high and can only explain up to around 18% of the variance. One potential explanation could be individual heterogeneity. We discuss this aspect in Section 1.4.1.

is held responsible. Models based on reciprocity and intentions can explain who is perceived responsible, but not when. We now test which motives have explanatory power when considering all motives simultaneously and, thereby, compare the importance of the different motives. Importantly, even though the outcome is the same in many scenarios in our experiment, the number of votes for and against the outcome differs and also which voter position was associated with which motive (i.e., whether the third, fourth, or fifth voter is pivotal; whether the first, second or third voter is the initiator). Table 1.2 shows the corresponding OLS regression outputs.

Table 1.2: Joint OLS regressions to compare the impact of the criteria on the usage of punishment and reward points

| | Punishment Points | | | Reward Points | |
|---|---|---|---|---|---|
| | Punishment | Both | | Reward | Both |
| Choice Unfair | 0.778*** | 0.466*** | Choice Fair | 0.455*** | 0.335*** |
| | (0.158) | (0.103) | | (0.111) | (0.084) |
| Outcome Unfair | -0.112 | -0.127* | Outcome Fair | -0.038 | -0.168** |
| | (0.111) | (0.052) | | (0.071) | (0.048) |
| Intention Unkind | -0.022 | 0.070 | Intention Kind | 0.180* | -0.001 |
| | (0.139) | (0.083) | | (0.082) | (0.042) |
| Initiator Unfair | 0.184 | 0.307 | Initiator Fair | 0.005 | -0.043 |
| | (0.228) | (0.242) | | (0.064) | (0.027) |
| Pivotal Unfair | 0.565** | 0.099 | Pivotal Fair | 0.213** | 0.052 |
| | (0.204) | (0.090) | | (0.076) | (0.046) |
| BF Responsibility (U) | 0.478 | 0.187 | BF Responsibility (F) | 0.285 | -0.233 |
| | (0.442) | (0.301) | | (0.348) | (0.167) |
| Ex-ante Engl Diff (U-F) | -0.104 | 0.004 | Ex-ante Engl Diff (F-U) | -0.048 | -0.006 |
| | (0.056) | (0.064) | | (0.042) | (0.029) |
| Ex-post Engl Diff (U-F) | 0.173* | 0.177*** | Ex-post Engl Diff (F-U) | 0.140 | 0.117** |
| | (0.076) | (0.044) | | (0.069) | (0.034) |
| Size of Majority | -0.019 | -0.059** | Size of Majority | -0.036 | 0.013 |
| | (0.026) | (0.021) | | (0.031) | (0.015) |
| Constant | 0.165 | 0.330** | Constant | 0.310* | 0.074 |
| | (0.127) | (0.097) | | (0.124) | (0.080) |
| Observations | 9,600 | 9,600 | Observations | 9,280 | 9,600 |
| R-squared | 0.210 | 0.152 | R-squared | 0.163 | 0.073 |
| Number of Subjects | 30 | 30 | Number of Subjects | 29 | 30 |

Note: OLS fixed effects regressions with punishment points and reward points as dependent variables. Punishment points (left side of the table) can take values from 0 to 7 and are used in the treatments *Punishment* and *Both*. Reward points (right side of the table) can take values from 0 to 7 and are used in the treatments *Reward* and *Both*. *Choice (Un)fair* equals 1 if the (un)fair allocation is chosen. *Outcome (Un)fair* is a dummy that equals 1 if the (un)fair outcome is implemented. *Intention (Un)kind* equals 1 if a voter votes for the (un)fair allocation while no majority was reached before. *Initiator (Un)fair* equals 1 if a voter is the initiator for the (un)fair outcome. *Pivotal (Un)fair* is an indicator that equals 1 if a voter is pivotal for the (un)fair outcome. *BF Responsibility (Un)fair* and *Ex-ante* and *Ex-post Engl Difference (U-F/F-U)* correspond to the responsibility measures explained in Section 1.3. *Size of Majority* indicates the number of majority voters and can take values from 3 to 5. Robust standard errors in parentheses clustered at the subject level.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

On the left side of the table we regress the punishment points on each sanctioning motive simultaneously for the treatments *Punishment* and *Both*, while the right side shows the respective regressions for the reward points in the treatments *Reward* and *Both*.[15] Since the measures are correlated, we also report regressions with only the positional variables (choice, outcome, intention, initiator and pivotal), with only the responsibility measures and with single regressors in Section 1.6.5 in the Appendix. We control for the size of the majority voters but do not include the voter position.[16]

The regression output in Table 1.2 shows that choices have an explanatory power on top of all other motives. Unfair choices predict the punishment patterns seen in the treatments *Punishment* and *Both*. In the *Punishment* treatment, voters who choose the unfair allocation get 0.778 more punishment points than voters choosing the fair allocation, and this increase equals 0.466 in the *Both* treatment. On the other side, fair choices are a good predictor for how people reward collective decisions in the treatments *Reward* and *Both*. In these treatments, voters are rewarded 0.455 / 0.335 more reward points when choosing the fair allocation compared to the unfair allocation. In the treatments *Punishment* and *Reward* pivotality has predictive power for the perceived responsibility when considering all motives. Being pivotal for the unfair outcome leads to 0.565 more punishment points, and being pivotal for the fair outcome leads to 0.213 more reward points compared to other intentional voters. Looking at the responsibility measures by Bartling and Fischbacher (2012) and Engl (2022) one can see that the criterion Ex-post Engl Difference for fair and unfair outcomes helps in explaining the punishment behavior in this joint regression. The remaining responsibility measures BF Responsibility and Ex-ante Engl Difference do not help much in explaining who is held responsible when combining all measures. This is because the responsibility measures encompass various individual sanction motives, which are included in this regression by other variables.[17] The results of the econometric comparison indicate that in sequential decisions, subjects mainly focus on the choices and the pivotal decision-maker when attributing responsibility.

We will now focus on the importance of the criterion Choice. First, we investigate what explains reward and punishment on top of Choice. The regressions are presented in Table A1.26 in the Appendix. On the one hand, they show that after a fair choice, there is almost no punishment and subjects do not differentiate between the different voters who voted for the fair outcome when punishing. The same is true for reward in case of an unfair choice. This result

[15] We report here the differences between the responsibility for unfair and fair outcomes for the Engl responsibility measures. The other variants can be found in Section 1.6.6 in the Appendix.

[16] We tested whether the voter position influences the perceived responsibility of each voter category by regressing the sanction points for each voter role on the voter position. The results show that only for the initiator and the pivotal voter, the positioning has an impact on the sanction points. Initiators on position 3 are punished more for bad outcomes than initiators on position 1 and 2. Initiators on position 2 and 3 are rewarded more for good outcomes than initiators on position 1. Pivotal voters on position 5 are punished more than pivotal voters on position 4 in unfair outcomes. Pivotal voters on position 5 are rewarded more than pivotal voters on position 3 and 4 in fair outcomes.

[17] The regressions in Section 1.6.5 of the Appendix show that the responsibility measures do not contribute much on top of the other variables, and their inclusion in the regression also does not much change the coefficients of the other variables. However, whether the positional variables are included or not affects explanatory power and coefficients of the responsibility measures. Their strength has to be assessed in the individual regressions, as discussed above.

can also be seen in Figure 1.3. On the other hand, the regressions show that among the voters with an unfair choice, the punishment is significantly explained by Pivotality. Again, the same is true for the reward among the voters with a fair choice. This suggests that the choice is a necessary condition for sanctioning. If the choice is fair, then there is no punishment and when the choice is unfair, there is no reward. Otherwise, more sophisticated criteria come into play.[18] Second, we explore whether subjects shift their strategies over time, especially since there are many decisions to take. For example, Choice is a relatively easy criterion and could become more pronounced towards the end of the experiment. This simpler (heuristic) strategy might come at the cost of other criteria, in particular Pivotality. The Tables A1.35 and A1.36 in the Appendix show the regression tables for decisions 1-21, 22-42, and 43-64. Choice does not lose predictive power over time, and Pivotality becomes even more important over time.

**Heterogeneity**

While the different theoretical models and motives have shown to be important in explaining the responsibility attribution pattern on average, the explanatory power of these measures is still quite low (see Figure 1.4). A potential reason for the low explanatory power of the joint analysis is heterogeneity in individual behavior. We perform finite mixture models to test if the overall data can be better explained by a mixture of different subgroups.[19] Using a bootstrap likelihood ratio test with 100 bootstrap replicates, we test if a model without subgroups (one component of coefficients) is a better fit to the data than a model with subgroups (more than one component of coefficients). In each treatment, the tests suggest that heterogeneity in individual behavior exists, as the data can be better explained by a mixture of more than one subgroup ($p < 0.02$ for all treatments).

Although there is only a limited number of subjects per treatment, we explore the individual patterns in each treatment. We identify the optimal number of clusters in each treatment by testing the best goodness of fit. Three different punishment and three different reward patterns are present in our experiment. Figure 1.5 shows the average punishment and reward points for the different voter categories used by recipients categorized within the same cluster.

Considering only punishment points, 30% of the subjects in the *Punishment* treatment use no or little punishment (Cluster 1).[20] However, most of the subjects in the *Punishment* treatment punish intentional voters and especially the pivotal voter for unfair outcomes and are categorized in Cluster 2. Subjects in Cluster 3 focus on punishing unfair choices. The cluster analysis in the *Reward* treatment shows that 25% in our experiment only use little reward (Cluster 1). Subjects in Cluster 2 focus on rewarding intentional voters for fair outcomes

[18] In Appendix 1.6.8, we present a hurdle model, in which we show what determines whether to sanction and what determines how much to sanction. However, this analysis does not provide new insights.

[19] We estimate a finite mixture model for each treatment (*Both* treatment separated by punishment and reward points) with the same positional variables as in Table 1.2. Varying the set of regressors to the full set of variables doesn't change the overall classification of components. The finite mixture models are estimated via a general linear regression using an EM-algorithm. We report the vector of coefficients for each subgroup across treatments in the Appendix 1.6.10.

[20] There are two subjects who do not use any punishment point across all decisions in the *Punishment* treatment.

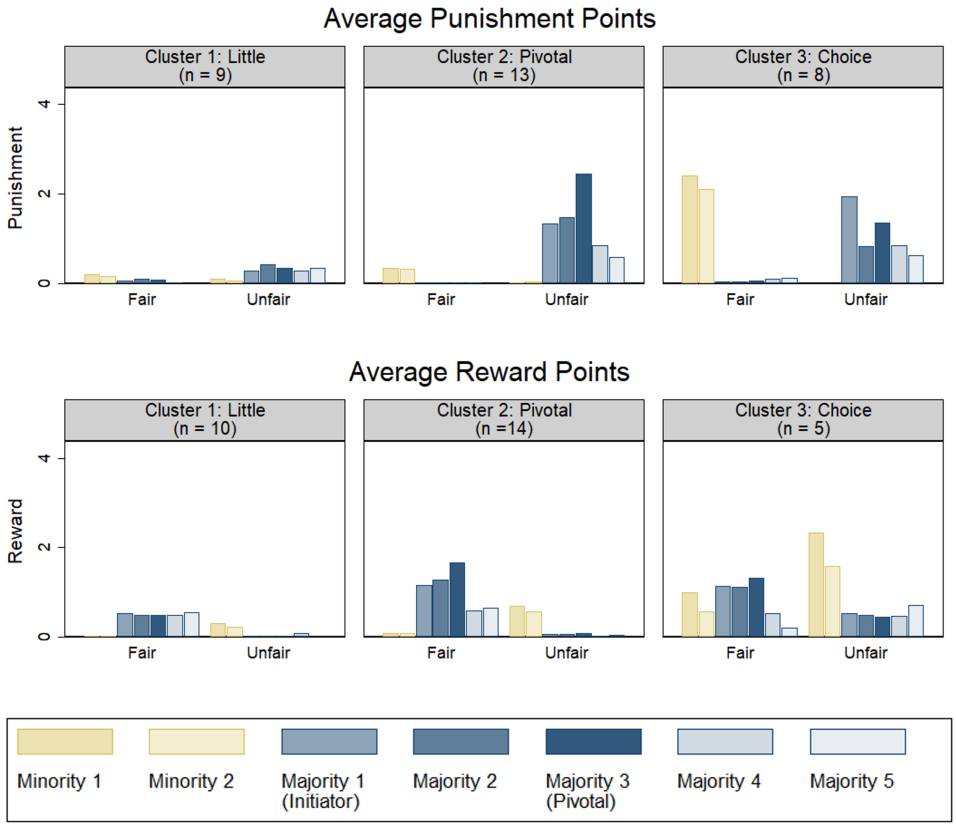


Figure 1.5: Cluster analysis: Punishment and reward patterns in *Punishment* and *Reward* treatment based on finite mixture models

Note: The figure shows the average punishment and reward points in absolute terms used in each cluster for fair and unfair outcomes across treatments. Hereby, the punishing patterns in the *Punishment* treatment are presented in the upper part of the figure, while the reward patterns in the *Reward* treatment are presented in the lower part of the figure. The number of subjects contained in each cluster per treatment are indicated in the titles of each sub-figure.

(especially the pivotal voter), while subjects categorized by Cluster 3 show a tendency to reward fair choices. Overall, the patterns shown in the treatments *Punishment* and *Reward* are very similar as these clusters can be described by: no or little punishment/reward, focus on the pivotal voter, and focus on choices.

The finite mixture models in the *Both* treatment are performed separately for punishment and reward points, and the results are shown in Figure 1.6. [21] The punishing and the rewarding behavior in the *Both* treatment can be described by three patterns: no punishment/reward, little sanctioning and sanctioning choices. These patterns are different than the ones observed in the *Punishment* and *Reward* treatment and suggest a less differentiated approach.

[21]Note that the magnitudes of average punishment and reward points shown in Figure 1.6 for each cluster in the *Both* treatment are smaller than the corresponding comparison in the *Punishment* and *Reward* treatments of Figure 1.5. This can be explained by our design, since in all treatments the maximum number of sanction points is limited to seven. This means that recipients in the *Both* treatment can use up to seven point for rewarding and punishing, while in the other two treatments the seven points can be used for the single sanction option. We added a minimal noise of absolute magnitude $\leq 0.01$ to the reward points when performing the finite mixture model, since the EM algorithm otherwise doesn't converge.

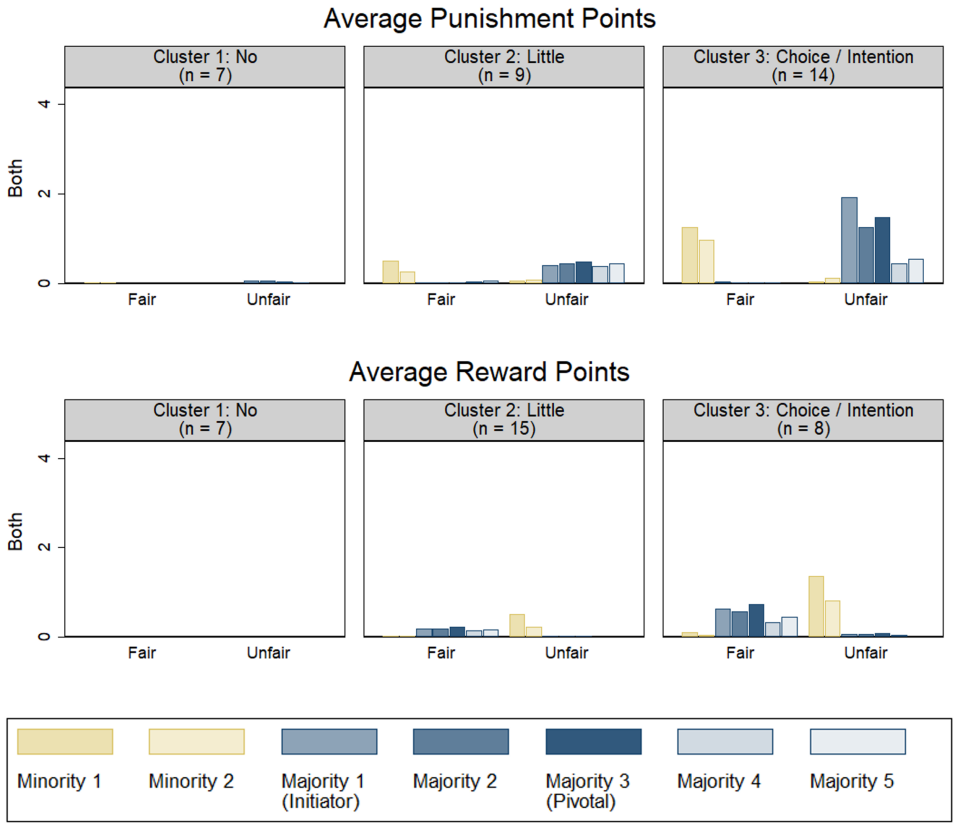


Figure 1.6: Cluster analysis: Punishment and reward patterns in the *Both* treatment based on finite mixture models

Note: The figure shows the average punishment and reward points in absolute terms used in each cluster for fair and unfair outcomes in the *Both* treatment. Hereby, the punishing patterns are presented in the upper part of the figure, while the reward patterns are presented in the lower part of the figure. The number of subjects contained in each cluster per treatment are indicated in titles of each sub-figure.

In sum, the presented exploratory cluster analysis indicates that there exist different types of responsibility attribution patterns. These results should be taken with care as there are only roughly 30 subjects per treatment and some clusters are rather small. Nonetheless, the patterns are interesting and suggest different types which are in line with different motives. We identify very similar punishment and reward patterns in our experiment. The expressed patterns are: little sanctioning, sanctioning according to intentions / pivotality and sanctioning according to choices. We find heterogeneity in individual behavior, which can be an explanation of the low predictive power of the theoretical models and motives in Table 1.2.

After having analyzed how recipients on an aggregate and individual level attribute responsibility, we complement the behavioral analysis by looking at the behavior of the voters.

### 1.4.2 Voting Behavior

We now turn to the voting behavior and analyze how subjects vote in collective decisions under the prospect of reward and punishment. Pivotality is an important aspect in the process

of responsibility attribution, as we have shown in the analysis of the recipients' behavior. Accordingly, strategic voters might take this into account and prevent (favor) being the target of punishment (reward) linked to pivotality.

We examine this strategic behavior by studying the voters' behavior in potentially pivotal decisions.[22] Table 1.3 offers a general overview of the share of decisions in which voters choose the unfair allocation depending on their position and the previous votes. We separate the results by our treatments (columns 3-5) and by the decisions of the voters in the two dictator decisions (columns 6-8). Importantly, we use the dictator game to elicit the preference of the voters for the fair or unfair allocation when neither a collective decision nor punishment or reward are implemented. 63 voters show a preference for the unfair allocation, while 14 voters show a fair preference. The remaining 13 voters have mixed preferences depending on the two allocation sets we offer them. Bold sequences in column 2 indicate situations in which voters face a potentially pivotal decision.

Table 1.3: Voting Behavior - Share of unfair choices

| | | By treatment | | | By preference in dictator games | | |
|---|---|---|---|---|---|---|---|
| Voter | Previous Voters | Both | Punishment | Reward | Fair (14 Voters) | Mixed (13 Voters) | Unfair (63 Voters) |
| 1 | - | 0.45 | 0.53 | 0.43 | 0.07 | 0.58 | 0.54 |
| 2 | U | 0.55 | 0.63 | 0.40 | 0.00 | 0.69 | 0.61 |
| | F | 0.48 | 0.47 | 0.42 | 0.04 | 0.54 | 0.53 |
| 3 | **UU** | 0.40 | 0.50 | 0.48 | 0.11 | 0.62 | 0.51 |
| | **FF** | 0.35 | 0.28 | 0.40 | 0.07 | 0.46 | 0.38 |
| | Tie | 0.48 | 0.60 | 0.49 | 0.11 | 0.60 | 0.60 |
| 4 | UUU | 0.23 | 0.35 | 0.40 | 0.18 | 0.54 | 0.32 |
| | FFF | 0.02 | 0.02 | 0.12 | 0.00 | 0.00 | 0.07 |
| | **2 of 3 U** | 0.42 | 0.52 | 0.53 | 0.07 | 0.46 | 0.59 |
| | **2 of 3 F** | 0.47 | 0.44 | 0.52 | 0.13 | 0.44 | 0.56 |
| 5 | 3 or more U | 0.31 | 0.30 | 0.40 | 0.26 | 0.52 | 0.31 |
| | 3 or more F | 0.05 | 0.03 | 0.14 | 0.02 | 0.05 | 0.09 |
| | **Tie** | 0.48 | 0.54 | 0.61 | 0.11 | 0.47 | 0.65 |

Note: Bold marked sequences in column 2 indicate decisions in which voters are potentially pivotal.

Voters who face a potentially pivotal decision are influenced by the choices of the previous voters. Across all treatments, the share of unfair choices for potentially pivotal voters is higher when the majority of previous voters voted unfair in comparison to a fair majority of previous votes (columns 3-5).[23] The biggest discrepancy results in the treatment *Punishment* for potentially pivotal voters on voting position three. 50% of these voters' decisions are unfair when

[22]Potentially pivotal means that exactly two of the previous voters voted for the same allocation, and that the own vote can be deterministic for the outcome. These situations can only appear for voters on positions three, four and five.

[23]One exception are the decisions of potentially pivotal voters in fourth position in the treatment *Both* where 47% of decisions are unfair following a fair majority, while only 42% are unfair following an unfair majority.

the first two voters voted unfair, while only 28% of the decisions are unfair when the first two voters voted fair.

The sequential decision design allows voters to use strategic (non-)delegation in order to avoid (seek) pivotality. Here, we focus on the potentially pivotal voters on either position three or four. In these cases, the voters can ensure being pivotal by following the majority of previous votes. But the voters can also vote against the majority of previous voters and can therefore delegate the notion of being pivotal to the next voter. Subjects showing a preference for the fair allocation in the dictator game mostly choose the fair allocation when being potentially pivotal (column 6 in Table 1.3). In contrast, voters expressing a preference for the unfair allocation in the dictator game often behave against their true preference in potentially pivotal situations of the collective decision (column 8). On voting positions three and four, potentially pivotal voters with an unfair preference ensure the unfair outcome by being pivotal in only 51% and 59% of all decisions, respectively. In 62% and 44% of the cases where the majority of previous voters chose the fair allocation, potentially pivotal voters on positions three and four voted against their true preference and decided on being pivotal for the fair outcome. On voting position 5 where no strategic (non-)delegation is possible, voters followed their true preference in only 65% of the cases.

Taken together, voters showing a preference for the unfair outcome often vote against their preference when being potentially pivotal. They avoid (seek) being pivotal for unfair (fair) outcomes by strategic (non-)delegation. But surprisingly many subjects do not try to appear fair when the outcome can no longer be changed, in particular when the outcome of the decision is unfair.

This completes our behavioral results. We now turn to the processing data we collected for recipients and voters to complement the behavioral analyses.

### 1.4.3 Process Measures

For the voters, we collected response time as a process measure. We hypothesize that response times inform us about the decision-making process, the difficulty of the decision and strategic decision-making (Konovalov and Ruff, 2022; Konovalov and Krajbich, 2019; Hausfeld et al., 2020; Spiliopoulos and Ortmann, 2018). A decision has to be considered as difficult if the voter strategically votes against the outcome preferences when being potentially pivotal. These decisions are characterized by a higher internal conflict, and should be accompanied by longer response times (Rubinstein, 2007). In Section 1.4.2, we showed that voters often decide against their true preference when being potentially pivotal. Potentially pivotal means that exactly two previous voters voted for the same allocation, and that the outcome can now be determined by the respective voter. This behavior suggests an internal conflict of being potentially pivotal.

Table 1.4 shows how the response time of voters is affected by being potentially pivotal, the voter position and the choice of the voters. When accounting for the voter position, we find that voters take significantly more time in choosing an allocation when they are potentially pivotal

(actual effect size is around 590 ms).[24] Another conflict shows when separating the voters by the true preference shown in the dictator game (last three columns of Table 1.4). First, being potentially pivotal still leads to higher response times for all types of voters. However, if voters vote against their true preferences, we find response times to be similarly affected. We find that fair types of voters take significantly more time when choosing the unfair allocation. In contrast, mixed and unfair types of voters spend less time when choosing the unfair allocation. Together with the results presented in Section 1.4.2, we can answer our fourth and fifth research question on how voters respond to the incentives created by responsibility attribution. Voters are aware of the responsibility that is linked to pivotality as they strategically use delegation to avoid punishment or non-delegation to gain reward even if it means that they vote against their true preference. This behavior is accompanied by a higher response time.

Table 1.4: Response Time Analyses Voters

| | Dependent Variable: Log Decision Time Voters | | | | | | |
|---|---|---|---|---|---|---|---|
| | Punishment | Reward | Both | Total | Fair Type | Unfair Type | Mixed Type |
| Potentially Pivotal | 0.164*** | 0.126** | 0.114** | 0.135*** | 0.107* | 0.154*** | 0.160** |
| | (0.031) | (0.040) | (0.036) | (0.021) | (0.036) | (0.028) | (0.042) |
| Voter Position | 0.037* | 0.010 | 0.038** | 0.028*** | 0.026 | 0.021 | 0.030 |
| | (0.014) | (0.013) | (0.012) | (0.008) | (0.018) | (0.011) | (0.018) |
| Choice Unfair | -0.095 | -0.135 | -0.102 | -0.115* | 0.252* | -0.165** | -0.156* |
| | (0.057) | (0.101) | (0.076) | (0.048) | (0.098) | (0.057) | (0.070) |
| Constant | 1.346*** | 1.397*** | 1.300*** | 1.351*** | 1.259*** | 1.379*** | 1.490*** |
| | (0.073) | (0.087) | (0.078) | (0.047) | (0.084) | (0.062) | (0.103) |
| Observations | 1,860 | 1,860 | 1,860 | 5,580 | 868 | 3,906 | 806 |
| R-squared | 0.032 | 0.019 | 0.023 | 0.024 | 0.029 | 0.033 | 0.036 |
| Number of Subjects | 30 | 30 | 30 | 90 | 14 | 63 | 13 |

Note: OLS regression with Decision Time of Voters (taken in log) as dependent variable. *Potentially Pivotal* is a dummy variable indicating whether the voter faces a decision where she can be pivotal depending on her choice. *Voter Position* indicates the position the voter holds in the decision she faces and can range from 1 to 5. *Choice Unfair* is a dummy variable indicating whether the voter chose the unfair or fair option.
Robust standard errors are clustered on individuals in parentheses, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

For the recipients, we collected gaze data as a process measure. We tracked the gaze of participants in order to analyze whether their sanctioning behavior is reflected in their information search. The gaze data reveals interesting insights about the importance of saliency, positioning of voters and the impact of sanctioning motives on the share of fixations. We present and discuss these analyses in Section 1.6.11 of the Appendix.

## 1.5 Conclusion

In our study, we use reward and punishment to investigate how people attribute responsibility in decision chains. In our experiment, five voters choose sequentially between two options of

[24]The average log decision time over all voting decisions is 1.48 which results in an average decision time of 4.41 seconds. Multiplying this average decision time by the coefficient of being potentially pivotal for the whole sample results in an actual effect size of 0.59 seconds.

how to allocate points between voters and recipients. One option is fair, the other is unfair. The recipients can reward and/or punish the voters, which we take as our measure of responsibility attribution. We test the relevance of different motives and measures and find that the actual choice, i.e., fair or unfair, plays a dominant role. It is clearly the most important determinant for the decision on whether to punish or not. Nevertheless, in line with Bartling et al. (2015), we find that the pivotal voter is assigned more punishment if the voters vote for an unfair allocation. We extend this result, showing that pivotality also matters when the voters vote for the fair allocation and the recipients can reward. Even though people have rather sophisticated responsibility attribution patterns, the conceptual models of Bartling and Fischbacher (2012) and Engl (2022) explained surprisingly little in comparison to, and in particular on top of the simple and mechanistic idea that it is the deed that determines responsibility. Overall, the explanatory power of the measures is not very high. This is partly due to heterogeneity. We show, using a finite mixture model, that three classes provide a better fit. Even though the sample size is too low to make strong statements about the distribution of types, in our sample there are subjects who sanction little, subjects who focus on choice and subjects who sanction the pivotal voter more than the others. Our results also show that the voters are aware of how responsibility is attributed and are particularly concerned when they are pivotal, which is also visible in longer response times.

Of course, specific features of the experiment could be relevant for the outcome, for example when investigating the role of the initiator. In our experiment, the two options were chosen with similar frequency. It is possible that when an action is rarely chosen, the initiator is assigned a higher level of responsibility. This could in particular be relevant in the case of fatalities. Another feature of our design could have reduced the relevance of the pivotal voter. Voters still had to vote, even when the decision was already made. Finishing the procedure when the result is determined could make the pivotal voter even more focal. Such variants could reveal the sensitivity of responsibility attribution to the specific situation. In addition, investigating costless reward and punishment could reveal whether selfish people apply different patterns of responsibility attribution. However, one can argue that also outside the lab, reward and punishment bear some cost, and therefore it is more important to know the responsibility attribution of people who are willing to bear such cost.

The responsibility attribution in our voting game captures situations, in which people take sequential decisions or actions that jointly generate an outcome. What does this imply in the real world, for example in the case of a disaster? First, any bad action is attributed some responsibility. Second, the pivotal person, i.e., the person after whose action the disaster was unavoidable, is generally assigned the highest responsibility. In his book, Whittingham (2004) observes that often the institutional environment is an important reason for disasters. Translated to our setting, he would consider the initiator as particularly responsible. However, we find few subjects who agree with this view. Of course, there are important differences to our lab experiments. First, in such disasters it is more difficult to identify the sequence. In particular, it is not easy to identify the pivotal agent - it is difficult to find out when the disaster

was no longer avoidable. Further, the different agents are less symmetric, both with respect to their contribution to the disaster and with respect to their formal responsibility. There are also many institutional details that could matter. Further research will allow investigating such variations, and our experiment provides a framework to do so.

# 1.6 Appendix

## 1.6.1 Supplemental Material: Experimental Design

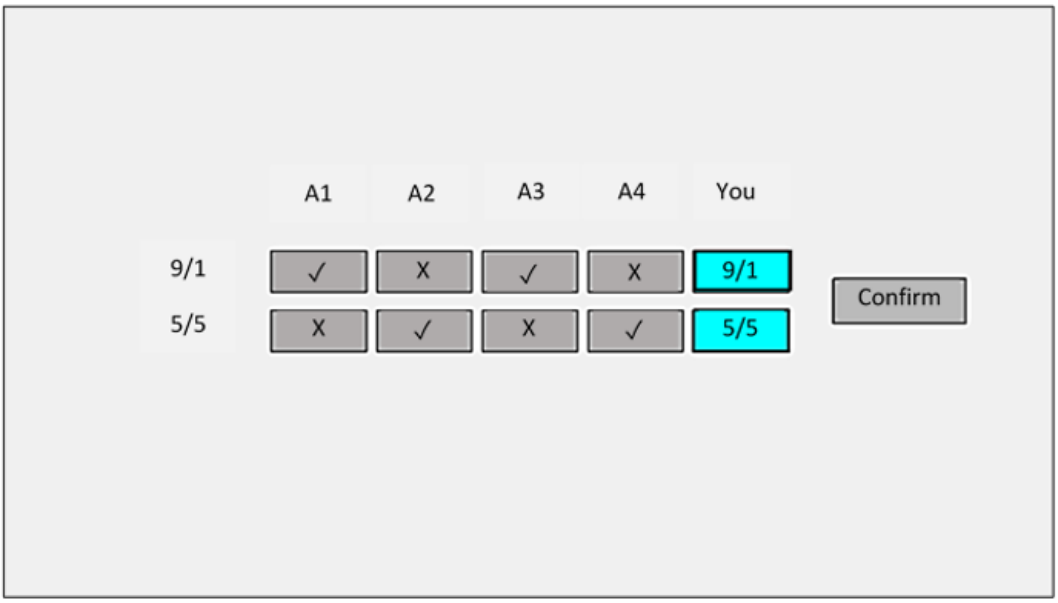


Figure A1.1: Exemplary screen for a voter

Note: The voter is on position five. Original text translated into English and font size enlarged for better readability.

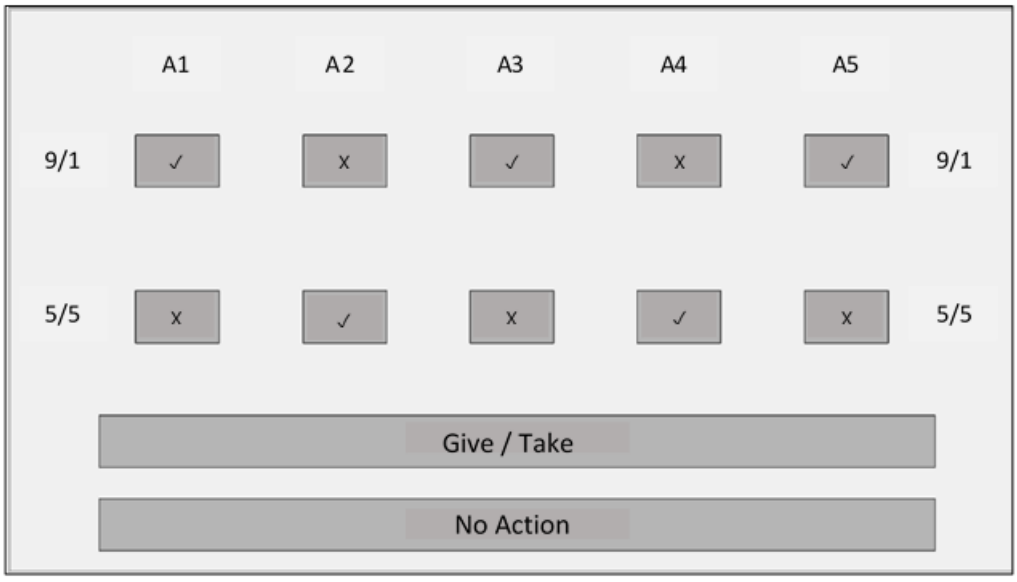


Figure A1.2: Exemplary first decision screen of a recipient

Note: The recipient is in the *Both* treatment. Original text translated into English and font size enlarged for better readability.

Table A1.1: Scenarios of Voters

| Voter 1 | Voter 2 | Voter 3 | Voter 4 | Voter 5 | Scenario | Voter Position |
|---|---|---|---|---|---|---|
| - | - | - | - | - | 1 | 1 |
| U | - | - | - | - | 2 | 2 |
| F | - | - | - | - | 3 | 2 |
| U | U | - | - | - | 4 | 3 |
| U | F | - | - | - | 5 | 3 |
| F | U | - | - | - | 6 | 3 |
| F | F | - | - | - | 7 | 3 |
| U | U | U | - | - | 8 | 4 |
| U | U | F | - | - | 9 | 4 |
| U | F | U | - | - | 10 | 4 |
| U | F | F | - | - | 11 | 4 |
| F | F | F | - | - | 12 | 4 |
| F | F | U | - | - | 13 | 4 |
| F | U | F | - | - | 14 | 4 |
| F | U | U | - | - | 15 | 4 |
| U | U | U | U | - | 16 | 5 |
| U | U | U | F | - | 17 | 5 |
| U | U | F | U | - | 18 | 5 |
| U | U | F | F | - | 19 | 5 |
| U | F | U | U | - | 20 | 5 |
| U | F | U | F | - | 21 | 5 |
| U | F | F | U | - | 22 | 5 |
| U | F | F | F | - | 23 | 5 |
| F | F | F | U | - | 24 | 5 |
| F | F | F | F | - | 25 | 5 |
| F | F | U | U | - | 26 | 5 |
| F | F | U | F | - | 27 | 5 |
| F | U | F | U | - | 28 | 5 |
| F | U | F | F | - | 29 | 5 |
| F | U | U | U | - | 30 | 5 |
| F | U | U | F | - | 31 | 5 |

## 1.6.2 Behavioral Results

Table A1.2: Average sanction points for different voter roles and voter positions

| | | Fair Outcome - Voter Position | | | | | | Unfair Outcome - Voter Position | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | Total | 1 | 2 | 3 | 4 | 5 | Total |
| Both | Minority 1 | -0.65 | -0.64 | -0.71 | -0.79 | -1.08 | -0.71 | 0.45 | 0.52 | 0.55 | 0.71 | 1.20 | 0.57 |
| | Minority 2 | . | -0.53 | -0.53 | -0.55 | -0.50 | -0.52 | . | 0.45 | 0.20 | 0.18 | 0.25 | 0.24 |
| | Initiator | 0.20 | 0.32 | 0.23 | . | . | 0.23 | -0.98 | -1.02 | -1.33 | . | . | -1.02 |
| | Majority 2 | . | 0.18 | 0.27 | 0.22 | . | 0.22 | . | -0.53 | -0.69 | -0.75 | . | -0.72 |
| | Pivotal | . | . | 0.27 | 0.22 | 0.41 | 0.30 | . | . | -0.65 | -0.73 | -1.02 | -0.82 |
| | Majority 4 | . | . | . | 0.17 | 0.12 | 0.14 | . | . | . | -0.37 | -0.29 | -0.32 |
| | Majority 5 | . | . | . | . | 0.18 | 0.18 | . | . | . | . | -0.38 | -0.38 |
| Punish | Minority 1 | -0.74 | -0.74 | -0.94 | -1.13 | -1.03 | -0.85 | -0.03 | -0.04 | -0.06 | -0.03 | -0.03 | -0.04 |
| | Minority 2 | . | -0.62 | -0.77 | -0.63 | -0.86 | -0.75 | . | -0.02 | -0.09 | 0 | -0.04 | -0.04 |
| | Initiator | -0.02 | -0.03 | -0.2 | . | . | -0.03 | -1.11 | -1.21 | -1.8 | . | . | -1.18 |
| | Majority 2 | . | 0 | -0.06 | -0.05 | . | -0.04 | . | -0.87 | -1.37 | -1.28 | . | -0.99 |
| | Pivotal | . | . | -0.06 | -0.06 | 0 | -0.04 | . | . | -1.54 | -1.4 | -1.64 | -1.52 |
| | Majority 4 | . | . | . | -0.05 | -0.03 | -0.04 | . | . | . | -0.63 | -0.7 | -0.67 |
| | Majority 5 | . | . | . | . | -0.03 | -0.03 | . | . | . | . | -0.52 | -0.52 |
| Reward | Minority 1 | 0.15 | 0.2 | 0.25 | 0.29 | 0.41 | 0.22 | 0.69 | 0.88 | 0.95 | 0.7 | 1.31 | 0.84 |
| | Minority 2 | . | 0.12 | 0.12 | 0.12 | 0.18 | 0.14 | . | 0.62 | 0.68 | 0.72 | 0.5 | 0.62 |
| | Initiator | 0.88 | 1.04 | 1.12 | . | . | 0.94 | 0.09 | 0.19 | 0.17 | . | . | 0.12 |
| | Majority 2 | . | 1.09 | 1.02 | 1.13 | . | 0.97 | . | 0.12 | 0.1 | 0.12 | . | 0.12 |
| | Pivotal | . | . | 1.01 | 1.14 | 1.36 | 1.19 | . | . | 0.09 | 0.12 | 0.16 | 0.13 |
| | Majority 4 | . | . | . | 0.54 | 0.54 | 0.54 | . | . | . | 0.12 | 0.12 | 0.12 |
| | Majority 5 | . | . | . | . | 0.53 | 0.53 | . | . | . | . | 0.14 | 0.14 |

Table A1.3: Average Sanction Points per Scenario

| | Both | | | | | Punishment | | | | | Reward | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | Voter 1 | Voter 2 | Voter 3 | Voter 4 | Voter 5 | Voter 1 | Voter 2 | Voter 3 | Voter 4 | Voter 5 | Voter 1 | Voter 2 | Voter 3 | Voter 4 | Voter 5 |
| U U U U U | -0.90 | -0.60 | -0.52 | -0.38 | -0.38 | -0.93 | -0.90 | -1.57 | -0.63 | -0.52 | 0.10 | 0.09 | 0.09 | 0.12 | 0.14 |
| U U U U F | -0.77 | -0.53 | -0.63 | -0.35 | 1.20 | -1.07 | -0.87 | -1.57 | -0.73 | -0.03 | 0.02 | 0.12 | 0.10 | 0.10 | 1.31 |
| U U U F U | -0.75 | -0.45 | -0.53 | 1.05 | -0.30 | -1.05 | -0.83 | -1.48 | -0.07 | -0.77 | 0.14 | 0.07 | 0.05 | 0.86 | 0.09 |
| U U U F F | -0.83 | -0.88 | -0.92 | 0.37 | 0.27 | -1.27 | -0.97 | -1.53 | 0.00 | -0.03 | 0.10 | 0.10 | 0.12 | 0.53 | 0.50 |
| U U F U U | -0.92 | -0.67 | 0.95 | -0.60 | -0.32 | -1.13 | -0.92 | -0.07 | -1.48 | -0.57 | 0.07 | 0.09 | 1.41 | 0.07 | 0.07 |
| U U F U F | -1.22 | -0.98 | 0.42 | -0.80 | 0.10 | -1.10 | -0.87 | -0.08 | -1.72 | 0.00 | 0.05 | 0.05 | 0.72 | 0.17 | 0.53 |
| U U F F U | -0.93 | -0.53 | 0.28 | 0.07 | -0.97 | -1.15 | -0.87 | -0.03 | 0.00 | -1.93 | 0.12 | 0.12 | 0.71 | 0.81 | 0.12 |
| U U F F F | -0.60 | -0.53 | 0.23 | 0.15 | 0.20 | -0.65 | -0.62 | -0.20 | -0.02 | -0.02 | 0.21 | 0.12 | 1.12 | 1.14 | 1.21 |
| U F U U U | -0.78 | 1.05 | -0.83 | -0.65 | -0.23 | -0.95 | -0.08 | -0.93 | -1.18 | -0.70 | 0.07 | 1.21 | 0.10 | 0.14 | 0.14 |
| U F U U F | -1.15 | 0.38 | -0.65 | -0.87 | 0.12 | -1.25 | 0.00 | -1.25 | -1.43 | -0.05 | 0.12 | 1.07 | 0.22 | 0.10 | 0.57 |
| U F U F U | -1.35 | 0.33 | -0.77 | 0.28 | -0.87 | -1.05 | 0.00 | -1.13 | 0.00 | -1.63 | 0.14 | 0.72 | 0.12 | 0.78 | 0.17 |
| U F U F F | -0.57 | 0.45 | -0.57 | 0.35 | 0.47 | -0.70 | 0.00 | -0.75 | -0.05 | 0.00 | 0.09 | 1.00 | 0.10 | 1.05 | 1.31 |
| U F F U U | -1.23 | 0.30 | 0.22 | -0.87 | -1.12 | -1.27 | -0.08 | -0.10 | -1.07 | -1.50 | 0.09 | 0.52 | 0.52 | 0.12 | 0.16 |
| U F F U F | -0.47 | 0.30 | 0.18 | -0.30 | 0.32 | -0.68 | 0.00 | -0.07 | -0.83 | 0.00 | 0.09 | 1.07 | 1.07 | 0.10 | 1.57 |
| U F F F U | -0.75 | 0.25 | 0.22 | 0.20 | -0.68 | -0.73 | -0.05 | -0.15 | -0.03 | -0.80 | 0.09 | 1.17 | 1.24 | 1.45 | 0.10 |
| U F F F F | -0.85 | 0.28 | 0.20 | 0.15 | 0.05 | -0.95 | -0.07 | -0.08 | -0.13 | -0.02 | 0.28 | 0.91 | 0.95 | 1.14 | 0.60 |
| F U U U U | 0.95 | -0.83 | -0.58 | -0.73 | -0.32 | -0.07 | -1.37 | -1.00 | -1.18 | -0.75 | 0.93 | 0.14 | 0.10 | 0.12 | 0.17 |
| F U U U F | 0.55 | -1.13 | -0.83 | -0.72 | 0.50 | 0.00 | -1.28 | -1.22 | -1.38 | -0.08 | 0.60 | 0.16 | 0.22 | 0.12 | 0.41 |
| F U U F U | 0.18 | -1.15 | -0.80 | 0.18 | -1.40 | 0.00 | -1.12 | -0.92 | 0.00 | -1.75 | 0.60 | 0.29 | 0.12 | 0.59 | 0.26 |
| F U U F F | 0.02 | -0.38 | -0.50 | 0.22 | 0.32 | -0.02 | -0.73 | -0.78 | 0.00 | 0.00 | 0.88 | 0.14 | 0.14 | 1.10 | 1.34 |
| F U F U U | 0.17 | -0.98 | 0.18 | -0.68 | -0.97 | -0.03 | -1.07 | -0.08 | -1.15 | -1.45 | 0.79 | 0.19 | 0.84 | 0.12 | 0.05 |
| F U F U F | 0.28 | -0.55 | 0.33 | -0.63 | 0.50 | 0.00 | -0.53 | -0.08 | -0.48 | 0.00 | 1.02 | 0.16 | 1.02 | 0.10 | 1.36 |
| F U F F U | 0.25 | -0.80 | 0.18 | 0.27 | -0.50 | 0.00 | -0.62 | -0.05 | -0.08 | -0.88 | 0.95 | 0.16 | 1.02 | 0.98 | 0.14 |
| F U F F F | 0.15 | -0.83 | 0.25 | 0.20 | 0.12 | -0.02 | -1.07 | -0.07 | -0.03 | 0.00 | 0.71 | 0.34 | 0.74 | 1.17 | 0.55 |
| F F U U U | 0.42 | 0.45 | -1.33 | -0.80 | -0.78 | -0.03 | -0.02 | -1.80 | -0.95 | -1.60 | 0.53 | 0.62 | 0.17 | 0.14 | 0.21 |
| F F U U F | 0.23 | 0.18 | -0.62 | -0.72 | 0.65 | 0.00 | 0.00 | -0.65 | -0.58 | 0.00 | 0.88 | 1.09 | 0.10 | 0.16 | 1.36 |
| F F U F U | 0.20 | 0.23 | -0.77 | 0.27 | -0.43 | 0.00 | 0.00 | -0.80 | -0.08 | -0.95 | 1.00 | 0.97 | 0.14 | 1.14 | 0.29 |
| F F U F F | 0.28 | 0.22 | -0.73 | 0.22 | 0.13 | 0.00 | -0.02 | -1.38 | 0.00 | -0.05 | 0.84 | 0.83 | 0.52 | 0.98 | 0.59 |
| F F F U U | 0.08 | 0.17 | 0.33 | -0.47 | -0.37 | 0.00 | -0.03 | -0.08 | -0.90 | -0.82 | 1.17 | 1.12 | 1.34 | 0.12 | 0.17 |
| F F F U F | 0.22 | 0.20 | 0.23 | -1.12 | 0.18 | -0.10 | -0.02 | -0.08 | -1.35 | -0.05 | 0.76 | 0.74 | 0.86 | 0.47 | 0.41 |
| F F F F U | 0.23 | 0.27 | 0.35 | 0.17 | -1.08 | 0.00 | 0.00 | -0.02 | -0.07 | -1.03 | 0.72 | 0.72 | 0.90 | 0.55 | 0.41 |
| F F F F F | 0.28 | 0.23 | 0.17 | 0.17 | 0.18 | -0.03 | -0.07 | -0.05 | -0.05 | -0.03 | 0.76 | 0.74 | 0.95 | 0.57 | 0.53 |

### 1.6.3 Responsibility Models

Table A1.4: Responsibility for voters in BF Model - Punishment Treatment (Allocation 9,1 vs. 5,5)

| | BF Responsibility (U) | | | | | BF Responsibility (F) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | Voter 1 | Voter 2 | Voter 3 | Voter 4 | Voter 5 | Voter 1 | Voter 2 | Voter 3 | Voter 4 | Voter 5 |
| UUUUU | 0.52 | 0.30 | 0.19 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUUUF | 0.52 | 0.30 | 0.19 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUUFU | 0.52 | 0.30 | 0.19 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUUFF | 0.52 | 0.30 | 0.19 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUFUU | 0.42 | 0.24 | 0.00 | 0.35 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUFUF | 0.42 | 0.24 | 0.00 | 0.35 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUFFU | 0.29 | 0.16 | 0.00 | 0.00 | 0.55 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.22 | 0.68 |
| UFUUU | 0.33 | 0.00 | 0.36 | 0.32 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UFUUF | 0.33 | 0.00 | 0.36 | 0.32 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UFUFU | 0.22 | 0.00 | 0.24 | 0.00 | 0.54 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UFUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.21 | 0.00 | 0.27 | 0.52 |
| UFFUU | 0.19 | 0.00 | 0.00 | 0.21 | 0.60 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UFFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.21 | 0.40 | 0.00 | 0.39 |
| UFFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.26 | 0.51 | 0.23 | 0.00 |
| UFFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.26 | 0.51 | 0.23 | 0.00 |
| FUUUU | 0.00 | 0.44 | 0.30 | 0.26 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FUUUF | 0.00 | 0.44 | 0.30 | 0.26 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FUUFU | 0.00 | 0.34 | 0.23 | 0.00 | 0.43 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FUUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.26 | 0.00 | 0.00 | 0.18 | 0.57 |
| FUFUU | 0.00 | 0.29 | 0.00 | 0.34 | 0.37 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FUFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.23 | 0.00 | 0.27 | 0.00 | 0.50 |
| FUFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.31 | 0.00 | 0.37 | 0.32 | 0.00 |
| FUFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.31 | 0.00 | 0.37 | 0.32 | 0.00 |
| FFUUU | 0.00 | 0.00 | 0.21 | 0.35 | 0.44 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FFUUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.25 | 0.22 | 0.00 | 0.00 | 0.53 |
| FFUFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.35 | 0.31 | 0.00 | 0.34 | 0.00 |
| FFUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.35 | 0.31 | 0.00 | 0.34 | 0.00 |
| FFFUU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.45 | 0.40 | 0.15 | 0.00 | 0.00 |
| FFFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.45 | 0.40 | 0.15 | 0.00 | 0.00 |
| FFFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.45 | 0.40 | 0.15 | 0.00 | 0.00 |
| FFFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.45 | 0.40 | 0.15 | 0.00 | 0.00 |

*Note:* The responsibility measure is the normalized version of raw responsibility.

Table A1.5: Responsibility for voters in BF Model - Punishment Treatment (Allocation 8,2 vs. 6,4)

| | BF Responsibility (U) | | | | | BF Responsibility (F) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | Voter 1 | Voter 2 | Voter 3 | Voter 4 | Voter 5 | Voter 1 | Voter 2 | Voter 3 | Voter 4 | Voter 5 |
| UUUUU | 0.42 | 0.33 | 0.25 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUUUF | 0.42 | 0.33 | 0.25 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUUFU | 0.42 | 0.33 | 0.25 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUUFF | 0.42 | 0.33 | 0.25 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUFUU | 0.35 | 0.27 | 0.00 | 0.37 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUFUF | 0.35 | 0.27 | 0.00 | 0.37 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUFFU | 0.26 | 0.20 | 0.00 | 0.00 | 0.54 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.12 | 0.27 | 0.61 |
| UFUUU | 0.28 | 0.00 | 0.36 | 0.36 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UFUUF | 0.28 | 0.00 | 0.36 | 0.36 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UFUFU | 0.21 | 0.00 | 0.26 | 0.00 | 0.53 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UFUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.26 | 0.00 | 0.29 | 0.44 |
| UFFUU | 0.21 | 0.00 | 0.00 | 0.26 | 0.53 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UFFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.27 | 0.29 | 0.00 | 0.45 |
| UFFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.37 | 0.40 | 0.23 | 0.00 |
| UFFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.37 | 0.40 | 0.23 | 0.00 |
| FUUUU | 0.00 | 0.34 | 0.29 | 0.37 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FUUUF | 0.00 | 0.34 | 0.29 | 0.37 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FUUFU | 0.00 | 0.27 | 0.23 | 0.00 | 0.51 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FUUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.21 | 0.00 | 0.00 | 0.22 | 0.57 |
| FUFUU | 0.00 | 0.24 | 0.00 | 0.31 | 0.45 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FUFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.19 | 0.00 | 0.30 | 0.00 | 0.51 |
| FUFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.27 | 0.00 | 0.43 | 0.30 | 0.00 |
| FUFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.27 | 0.00 | 0.43 | 0.30 | 0.00 |
| FFUUU | 0.00 | 0.00 | 0.22 | 0.32 | 0.46 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FFUUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.20 | 0.22 | 0.00 | 0.00 | 0.58 |
| FFUFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.29 | 0.32 | 0.00 | 0.39 | 0.00 |
| FFUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.29 | 0.32 | 0.00 | 0.39 | 0.00 |
| FFFUU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.41 | 0.46 | 0.13 | 0.00 | 0.00 |
| FFFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.41 | 0.46 | 0.13 | 0.00 | 0.00 |
| FFFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.41 | 0.46 | 0.13 | 0.00 | 0.00 |
| FFFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.41 | 0.46 | 0.13 | 0.00 | 0.00 |

Table A1.6: Responsibility for voters in BF Model - Reward Treatment (Allocation 9,1 vs. 5,5)

| | BF Responsibility (U) | | | | | BF Responsibility (F) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | Voter 1 | Voter 2 | Voter 3 | Voter 4 | Voter 5 | Voter 1 | Voter 2 | Voter 3 | Voter 4 | Voter 5 |
| UUUUU | 0.32 | 0.46 | 0.22 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUUUF | 0.32 | 0.46 | 0.22 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUUFU | 0.32 | 0.46 | 0.22 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUUFF | 0.32 | 0.46 | 0.22 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUFUU | 0.26 | 0.38 | 0.00 | 0.35 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUFUF | 0.26 | 0.38 | 0.00 | 0.35 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUFFU | 0.19 | 0.28 | 0.00 | 0.00 | 0.52 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.11 | 0.22 | 0.67 |
| UFUUU | 0.25 | 0.00 | 0.45 | 0.29 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UFUUF | 0.25 | 0.00 | 0.45 | 0.29 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UFUFU | 0.18 | 0.00 | 0.33 | 0.00 | 0.49 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UFUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.13 | 0.00 | 0.24 | 0.63 |
| UFFUU | 0.18 | 0.00 | 0.00 | 0.45 | 0.36 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UFFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.12 | 0.22 | 0.00 | 0.66 |
| UFFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.18 | 0.34 | 0.48 | 0.00 |
| UFFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.18 | 0.34 | 0.48 | 0.00 |
| FUUUU | 0.00 | 0.43 | 0.34 | 0.23 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FUUUF | 0.00 | 0.43 | 0.34 | 0.23 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FUUFU | 0.00 | 0.34 | 0.27 | 0.00 | 0.39 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FUUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.15 | 0.00 | 0.00 | 0.16 | 0.69 |
| FUFUU | 0.00 | 0.31 | 0.00 | 0.37 | 0.31 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FUFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.13 | 0.00 | 0.21 | 0.00 | 0.66 |
| FUFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.19 | 0.00 | 0.30 | 0.52 | 0.00 |
| FUFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.19 | 0.00 | 0.30 | 0.52 | 0.00 |
| FFUUU | 0.00 | 0.00 | 0.24 | 0.37 | 0.39 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FFUUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.14 | 0.22 | 0.00 | 0.00 | 0.64 |
| FFUFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.20 | 0.32 | 0.00 | 0.48 | 0.00 |
| FFUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.20 | 0.32 | 0.00 | 0.48 | 0.00 |
| FFFUU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.28 | 0.43 | 0.29 | 0.00 | 0.00 |
| FFFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.28 | 0.43 | 0.29 | 0.00 | 0.00 |
| FFFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.28 | 0.43 | 0.29 | 0.00 | 0.00 |
| FFFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.28 | 0.43 | 0.29 | 0.00 | 0.00 |

Table A1.7: Responsibility for voters in BF Model - Reward Treatment (Allocation 8,2 vs. 6,4)

| | BF Responsibility (U) | | | | | BF Responsibility (F) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | Voter 1 | Voter 2 | Voter 3 | Voter 4 | Voter 5 | Voter 1 | Voter 2 | Voter 3 | Voter 4 | Voter 5 |
| UUUUU | 0.40 | 0.42 | 0.18 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUUUF | 0.40 | 0.42 | 0.18 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUUFU | 0.40 | 0.42 | 0.18 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUUFF | 0.40 | 0.42 | 0.18 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUFUU | 0.34 | 0.36 | 0.00 | 0.30 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUFUF | 0.34 | 0.36 | 0.00 | 0.30 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUFFU | 0.25 | 0.26 | 0.00 | 0.00 | 0.49 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.11 | 0.30 | 0.60 |
| UFUUU | 0.30 | 0.00 | 0.26 | 0.44 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UFUUF | 0.30 | 0.00 | 0.26 | 0.44 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UFUFU | 0.23 | 0.00 | 0.20 | 0.00 | 0.57 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UFUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.22 | 0.00 | 0.27 | 0.51 |
| UFFUU | 0.25 | 0.00 | 0.00 | 0.30 | 0.46 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UFFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.20 | 0.19 | 0.00 | 0.61 |
| UFFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.29 | 0.27 | 0.44 | 0.00 |
| UFFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.29 | 0.27 | 0.44 | 0.00 |
| FUUUU | 0.00 | 0.40 | 0.40 | 0.20 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FUUUF | 0.00 | 0.40 | 0.40 | 0.20 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FUUFU | 0.00 | 0.30 | 0.29 | 0.00 | 0.41 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FUUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.14 | 0.00 | 0.00 | 0.26 | 0.61 |
| FUFUU | 0.00 | 0.29 | 0.00 | 0.28 | 0.43 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FUFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.14 | 0.00 | 0.29 | 0.00 | 0.57 |
| FUFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.19 | 0.00 | 0.41 | 0.40 | 0.00 |
| FUFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.19 | 0.00 | 0.41 | 0.40 | 0.00 |
| FFUUU | 0.00 | 0.00 | 0.22 | 0.30 | 0.49 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FFUUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.16 | 0.19 | 0.00 | 0.00 | 0.65 |
| FFUFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.22 | 0.28 | 0.00 | 0.50 | 0.00 |
| FFUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.22 | 0.28 | 0.00 | 0.50 | 0.00 |
| FFFUU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.33 | 0.41 | 0.27 | 0.00 | 0.00 |
| FFFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.33 | 0.41 | 0.27 | 0.00 | 0.00 |
| FFFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.33 | 0.41 | 0.27 | 0.00 | 0.00 |
| FFFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.33 | 0.41 | 0.27 | 0.00 | 0.00 |

Table A1.8: Responsibility for voters in BF Model - Both Treatment (Allocation 9,1 vs. 5,5)

| | BF Responsibility (U) | | | | | BF Responsibility (F) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | Voter 1 | Voter 2 | Voter 3 | Voter 4 | Voter 5 | Voter 1 | Voter 2 | Voter 3 | Voter 4 | Voter 5 |
| UUUUU | 0.36 | 0.34 | 0.31 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUUUF | 0.36 | 0.34 | 0.31 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUUFU | 0.36 | 0.34 | 0.31 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUUFF | 0.36 | 0.34 | 0.31 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUFUU | 0.30 | 0.28 | 0.00 | 0.42 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUFUF | 0.30 | 0.28 | 0.00 | 0.42 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUFFU | 0.24 | 0.23 | 0.00 | 0.00 | 0.54 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.14 | 0.21 | 0.65 |
| UFUUU | 0.27 | 0.00 | 0.41 | 0.32 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UFUUF | 0.27 | 0.00 | 0.41 | 0.32 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UFUFU | 0.21 | 0.00 | 0.32 | 0.00 | 0.47 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UFUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.21 | 0.00 | 0.23 | 0.56 |
| UFFUU | 0.20 | 0.00 | 0.00 | 0.24 | 0.56 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UFFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.22 | 0.28 | 0.00 | 0.50 |
| UFFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.31 | 0.39 | 0.30 | 0.00 |
| UFFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.31 | 0.39 | 0.30 | 0.00 |
| FUUUU | 0.00 | 0.34 | 0.30 | 0.36 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FUUUF | 0.00 | 0.34 | 0.30 | 0.36 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FUUFU | 0.00 | 0.27 | 0.24 | 0.00 | 0.49 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FUUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.16 | 0.00 | 0.00 | 0.20 | 0.65 |
| FUFUU | 0.00 | 0.25 | 0.00 | 0.23 | 0.52 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FUFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.16 | 0.00 | 0.28 | 0.00 | 0.57 |
| FUFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.21 | 0.00 | 0.37 | 0.43 | 0.00 |
| FUFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.21 | 0.00 | 0.37 | 0.43 | 0.00 |
| FFUUU | 0.00 | 0.00 | 0.15 | 0.34 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FFUUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.16 | 0.24 | 0.00 | 0.00 | 0.60 |
| FFUFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.24 | 0.38 | 0.00 | 0.38 | 0.00 |
| FFUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.24 | 0.38 | 0.00 | 0.38 | 0.00 |
| FFFUU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.33 | 0.51 | 0.17 | 0.00 | 0.00 |
| FFFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.33 | 0.51 | 0.17 | 0.00 | 0.00 |
| FFFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.33 | 0.51 | 0.17 | 0.00 | 0.00 |
| FFFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.33 | 0.51 | 0.17 | 0.00 | 0.00 |

Table A1.9: Responsibility for voters in BF Model - Both Treatment (Allocation 8,2 vs. 6,4)

| | BF Responsibility (U) | | | | | BF Responsibility (F) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | Voter 1 | Voter 2 | Voter 3 | Voter 4 | Voter 5 | Voter 1 | Voter 2 | Voter 3 | Voter 4 | Voter 5 |
| UUUUU | 0.33 | 0.24 | 0.43 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUUUF | 0.33 | 0.24 | 0.43 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUUFU | 0.33 | 0.24 | 0.43 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUUFF | 0.33 | 0.24 | 0.43 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUFUU | 0.25 | 0.19 | 0.00 | 0.56 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUFUF | 0.25 | 0.19 | 0.00 | 0.56 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUFFU | 0.19 | 0.14 | 0.00 | 0.00 | 0.67 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UUFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.21 | 0.31 | 0.48 |
| UFUUU | 0.24 | 0.00 | 0.34 | 0.42 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UFUUF | 0.24 | 0.00 | 0.34 | 0.42 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UFUFU | 0.18 | 0.00 | 0.25 | 0.00 | 0.56 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UFUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.23 | 0.00 | 0.28 | 0.49 |
| UFFUU | 0.19 | 0.00 | 0.00 | 0.24 | 0.58 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| UFFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.23 | 0.26 | 0.00 | 0.51 |
| UFFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.32 | 0.36 | 0.32 | 0.00 |
| UFFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.32 | 0.36 | 0.32 | 0.00 |
| FUUUU | 0.00 | 0.24 | 0.37 | 0.38 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FUUUF | 0.00 | 0.24 | 0.37 | 0.38 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FUUFU | 0.00 | 0.18 | 0.28 | 0.00 | 0.54 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FUUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.20 | 0.00 | 0.00 | 0.28 | 0.52 |
| FUFUU | 0.00 | 0.18 | 0.00 | 0.20 | 0.61 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FUFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.22 | 0.00 | 0.29 | 0.00 | 0.49 |
| FUFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.29 | 0.00 | 0.39 | 0.32 | 0.00 |
| FUFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.29 | 0.00 | 0.39 | 0.32 | 0.00 |
| FFUUU | 0.00 | 0.00 | 0.15 | 0.31 | 0.54 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FFUUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.21 | 0.21 | 0.00 | 0.00 | 0.58 |
| FFUFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.32 | 0.31 | 0.00 | 0.37 | 0.00 |
| FFUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.32 | 0.31 | 0.00 | 0.37 | 0.00 |
| FFFUU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.41 | 0.41 | 0.18 | 0.00 | 0.00 |
| FFFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.41 | 0.41 | 0.18 | 0.00 | 0.00 |
| FFFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.41 | 0.41 | 0.18 | 0.00 | 0.00 |
| FFFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.41 | 0.41 | 0.18 | 0.00 | 0.00 |

Table A1.10: Causal Responsibility for voters in simple variant of Engl Model - Punishment Treatment (Allocation 9,1 vs. 5,5)

| | Ex-post Responsibility (U) | | | | | Ex-ante Responsibility (U) | | | | | Ex-post Responsibility (F) | | | | | Ex-ante Responsibility (F) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | V1 | V2 | V3 | V4 | V5 | V1 | V2 | V3 | V4 | V5 | V1 | V2 | V3 | V4 | V5 | V1 | V2 | V3 | V4 | V5 |
| UUUUU | 0.65 | 0.40 | 0.16 | 0.00 | 0.00 | 0.54 | 0.37 | 0.16 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.04 | 0.00 | 0.00 | 0.00 |
| UUUUF | 0.65 | 0.40 | 0.16 | 0.00 | 0.00 | 0.54 | 0.37 | 0.16 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.04 | 0.00 | 0.00 | 0.00 |
| UUUFU | 0.65 | 0.40 | 0.16 | 0.00 | 0.00 | 0.54 | 0.37 | 0.16 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.04 | 0.00 | 0.00 | 0.00 |
| UUUFF | 0.65 | 0.40 | 0.16 | 0.00 | 0.00 | 0.54 | 0.37 | 0.16 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.04 | 0.00 | 0.00 | 0.00 |
| UUFUU | 0.65 | 0.40 | 0.00 | 0.37 | 0.00 | 0.54 | 0.37 | 0.00 | 0.37 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.04 | 0.16 | 0.00 | 0.00 |
| UUFUF | 0.65 | 0.40 | 0.00 | 0.37 | 0.00 | 0.54 | 0.37 | 0.00 | 0.37 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.04 | 0.16 | 0.00 | 0.00 |
| UUFFU | 0.65 | 0.40 | 0.00 | 0.00 | 1.00 | 0.54 | 0.37 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.04 | 0.16 | 0.37 | 0.00 |
| UUFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.54 | 0.37 | 0.00 | 0.00 | 0.00 | 0.35 | 0.60 | 1.00 | 1.00 | 1.00 | 0.06 | 0.04 | 0.16 | 0.37 | 1.00 |
| UFUUU | 0.65 | 0.07 | 0.81 | 0.47 | 0.00 | 0.54 | 0.04 | 0.66 | 0.47 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.37 | 0.03 | 0.00 | 0.00 |
| UFUUF | 0.65 | 0.07 | 0.81 | 0.47 | 0.00 | 0.54 | 0.04 | 0.66 | 0.47 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.37 | 0.03 | 0.00 | 0.00 |
| UFUFU | 0.65 | 0.07 | 0.81 | 0.00 | 1.00 | 0.54 | 0.04 | 0.66 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.37 | 0.03 | 0.47 | 0.00 |
| UFUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.54 | 0.04 | 0.66 | 0.00 | 0.00 | 0.35 | 0.93 | 0.19 | 1.00 | 1.00 | 0.06 | 0.37 | 0.03 | 0.47 | 1.00 |
| UFFUU | 0.65 | 0.07 | 0.19 | 1.00 | 1.00 | 0.54 | 0.04 | 0.03 | 0.40 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.37 | 0.66 | 0.00 | 0.00 |
| UFFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.54 | 0.04 | 0.03 | 0.40 | 0.00 | 0.35 | 0.93 | 0.81 | 0.00 | 1.00 | 0.06 | 0.37 | 0.66 | 0.00 | 1.00 |
| UFFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.54 | 0.04 | 0.03 | 0.00 | 0.00 | 0.35 | 0.93 | 0.81 | 0.40 | 0.00 | 0.06 | 0.37 | 0.66 | 0.40 | 0.00 |
| UFFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.54 | 0.04 | 0.03 | 0.00 | 0.00 | 0.35 | 0.93 | 0.81 | 0.40 | 0.00 | 0.06 | 0.37 | 0.66 | 0.40 | 0.00 |
| FUUUU | 0.18 | 0.91 | 0.70 | 0.37 | 0.00 | 0.05 | 0.57 | 0.58 | 0.37 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.54 | 0.03 | 0.05 | 0.00 | 0.00 |
| FUUUF | 0.18 | 0.91 | 0.70 | 0.37 | 0.00 | 0.05 | 0.57 | 0.58 | 0.37 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.54 | 0.03 | 0.05 | 0.00 | 0.00 |
| FUUFU | 0.18 | 0.91 | 0.70 | 0.00 | 1.00 | 0.05 | 0.57 | 0.58 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.54 | 0.03 | 0.05 | 0.37 | 0.00 |
| FUUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.57 | 0.58 | 0.00 | 0.00 | 0.82 | 0.09 | 0.30 | 1.00 | 1.00 | 0.54 | 0.03 | 0.05 | 0.37 | 1.00 |
| FUFUU | 0.18 | 0.91 | 0.17 | 1.00 | 1.00 | 0.05 | 0.57 | 0.05 | 0.63 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.54 | 0.03 | 0.58 | 0.00 | 0.00 |
| FUFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.57 | 0.05 | 0.63 | 0.00 | 0.82 | 0.09 | 0.83 | 0.00 | 1.00 | 0.54 | 0.03 | 0.58 | 0.00 | 1.00 |
| FUFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.57 | 0.05 | 0.00 | 0.00 | 0.82 | 0.09 | 0.83 | 0.63 | 0.00 | 0.54 | 0.03 | 0.58 | 0.63 | 0.00 |
| FUFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.57 | 0.05 | 0.00 | 0.00 | 0.82 | 0.09 | 0.83 | 0.63 | 0.00 | 0.54 | 0.03 | 0.58 | 0.63 | 0.00 |
| FFUUU | 0.18 | 0.37 | 1.00 | 1.00 | 1.00 | 0.05 | 0.03 | 0.28 | 0.60 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.54 | 0.57 | 0.00 | 0.00 | 0.00 |
| FFUUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.03 | 0.28 | 0.60 | 0.00 | 0.82 | 0.63 | 0.00 | 0.00 | 1.00 | 0.54 | 0.57 | 0.00 | 0.00 | 1.00 |
| FFUFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.03 | 0.28 | 0.00 | 0.00 | 0.82 | 0.63 | 0.00 | 0.00 | 0.00 | 0.54 | 0.57 | 0.00 | 0.60 | 0.00 |
| FFUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.03 | 0.28 | 0.00 | 0.00 | 0.82 | 0.63 | 0.00 | 0.60 | 0.00 | 0.54 | 0.57 | 0.00 | 0.60 | 0.00 |
| FFFUU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.03 | 0.00 | 0.00 | 0.00 | 0.82 | 0.63 | 0.28 | 0.00 | 0.00 | 0.54 | 0.57 | 0.28 | 0.00 | 0.00 |
| FFFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.03 | 0.00 | 0.00 | 0.00 | 0.82 | 0.63 | 0.28 | 0.00 | 0.00 | 0.54 | 0.57 | 0.28 | 0.00 | 0.00 |
| FFFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.03 | 0.00 | 0.00 | 0.00 | 0.82 | 0.63 | 0.28 | 0.00 | 0.00 | 0.54 | 0.57 | 0.28 | 0.00 | 0.00 |
| FFFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.03 | 0.00 | 0.00 | 0.00 | 0.82 | 0.63 | 0.28 | 0.00 | 0.00 | 0.54 | 0.57 | 0.28 | 0.00 | 0.00 |

Table A1.11: Causal Responsibility for voters in simple variant of Engl Model - Punishment Treatment (Allocation 8,2 vs. 6,4)

| | Ex-post Responsibility (U) | | | | | Ex-ante Responsibility (U) | | | | | Ex-post Responsibility (F) | | | | | Ex-ante Responsibility (F) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | V1 | V2 | V3 | V4 | V5 | V1 | V2 | V3 | V4 | V5 | V1 | V2 | V3 | V4 | V5 | V1 | V2 | V3 | V4 | V5 |
| UUUUU | 0.72 | 0.56 | 0.23 | 0.00 | 0.00 | 0.50 | 0.49 | 0.23 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.06 | 0.00 | 0.00 | 0.00 |
| UUUUF | 0.72 | 0.56 | 0.23 | 0.00 | 0.00 | 0.50 | 0.49 | 0.23 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.06 | 0.00 | 0.00 | 0.00 |
| UUUFU | 0.72 | 0.56 | 0.23 | 0.00 | 0.00 | 0.50 | 0.49 | 0.23 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.06 | 0.00 | 0.00 | 0.00 |
| UUUFF | 0.72 | 0.56 | 0.23 | 0.00 | 0.00 | 0.50 | 0.49 | 0.23 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.06 | 0.00 | 0.00 | 0.00 |
| UUFUU | 0.72 | 0.56 | 0.00 | 0.47 | 0.00 | 0.50 | 0.49 | 0.00 | 0.47 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.06 | 0.23 | 0.00 | 0.00 |
| UUFUF | 0.72 | 0.56 | 0.00 | 0.47 | 0.00 | 0.50 | 0.49 | 0.00 | 0.47 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.06 | 0.23 | 0.00 | 0.00 |
| UUFFU | 0.72 | 0.56 | 0.00 | 0.00 | 1.00 | 0.50 | 0.49 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.06 | 0.23 | 0.47 | 0.00 |
| UUFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.50 | 0.49 | 0.00 | 0.00 | 0.00 | 0.28 | 0.44 | 1.00 | 1.00 | 1.00 | 0.08 | 0.06 | 0.23 | 0.47 | 1.00 |
| UFUUU | 0.72 | 0.13 | 0.84 | 0.57 | 0.00 | 0.50 | 0.06 | 0.60 | 0.57 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.49 | 0.05 | 0.00 | 0.00 |
| UFUUF | 0.72 | 0.13 | 0.84 | 0.57 | 0.00 | 0.50 | 0.06 | 0.60 | 0.57 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.49 | 0.05 | 0.00 | 0.00 |
| UFUFU | 0.72 | 0.13 | 0.84 | 0.00 | 1.00 | 0.50 | 0.06 | 0.60 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.49 | 0.05 | 0.57 | 0.00 |
| UFUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.50 | 0.06 | 0.60 | 0.00 | 0.00 | 0.28 | 0.87 | 0.16 | 1.00 | 1.00 | 0.08 | 0.49 | 0.05 | 0.57 | 1.00 |
| UFFUU | 0.72 | 0.13 | 0.28 | 1.00 | 1.00 | 0.50 | 0.06 | 0.05 | 0.43 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.49 | 0.60 | 0.00 | 0.00 |
| UFFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.50 | 0.06 | 0.05 | 0.43 | 0.00 | 0.28 | 0.87 | 0.72 | 0.00 | 1.00 | 0.08 | 0.49 | 0.60 | 0.00 | 1.00 |
| UFFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.50 | 0.06 | 0.05 | 0.00 | 0.00 | 0.28 | 0.87 | 0.72 | 0.43 | 0.00 | 0.08 | 0.49 | 0.60 | 0.43 | 0.00 |
| UFFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.50 | 0.06 | 0.05 | 0.00 | 0.00 | 0.28 | 0.87 | 0.72 | 0.43 | 0.00 | 0.08 | 0.49 | 0.60 | 0.43 | 0.00 |
| FUUUU | 0.30 | 0.94 | 0.79 | 0.47 | 0.00 | 0.06 | 0.49 | 0.58 | 0.47 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.50 | 0.03 | 0.06 | 0.00 | 0.00 |
| FUUUF | 0.30 | 0.94 | 0.79 | 0.47 | 0.00 | 0.06 | 0.49 | 0.58 | 0.47 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.50 | 0.03 | 0.06 | 0.00 | 0.00 |
| FUUFU | 0.30 | 0.94 | 0.79 | 0.00 | 1.00 | 0.06 | 0.49 | 0.58 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.50 | 0.03 | 0.06 | 0.47 | 0.00 |
| FUUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.49 | 0.58 | 0.00 | 0.00 | 0.70 | 0.06 | 0.21 | 1.00 | 1.00 | 0.50 | 0.03 | 0.06 | 0.47 | 1.00 |
| FUFUU | 0.30 | 0.94 | 0.26 | 1.00 | 1.00 | 0.06 | 0.49 | 0.06 | 0.53 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.50 | 0.03 | 0.58 | 0.00 | 0.00 |
| FUFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.49 | 0.06 | 0.53 | 0.00 | 0.70 | 0.06 | 0.74 | 0.00 | 1.00 | 0.50 | 0.03 | 0.58 | 0.00 | 1.00 |
| FUFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.49 | 0.06 | 0.00 | 0.00 | 0.70 | 0.06 | 0.74 | 0.53 | 0.00 | 0.50 | 0.03 | 0.58 | 0.53 | 0.00 |
| FUFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.49 | 0.06 | 0.00 | 0.00 | 0.70 | 0.06 | 0.74 | 0.53 | 0.00 | 0.50 | 0.03 | 0.58 | 0.53 | 0.00 |
| FFUUU | 0.30 | 0.47 | 1.00 | 1.00 | 1.00 | 0.06 | 0.03 | 0.26 | 0.57 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.50 | 0.49 | 0.00 | 0.00 | 0.00 |
| FFUUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.03 | 0.26 | 0.57 | 0.00 | 0.70 | 0.53 | 0.00 | 0.00 | 1.00 | 0.50 | 0.49 | 0.00 | 0.00 | 1.00 |
| FFUFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.03 | 0.26 | 0.00 | 0.00 | 0.70 | 0.53 | 0.00 | 0.00 | 0.00 | 0.50 | 0.49 | 0.00 | 0.57 | 0.00 |
| FFUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.03 | 0.26 | 0.00 | 0.00 | 0.70 | 0.53 | 0.00 | 0.57 | 0.00 | 0.50 | 0.49 | 0.00 | 0.57 | 0.00 |
| FFFUU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.03 | 0.00 | 0.00 | 0.00 | 0.70 | 0.53 | 0.26 | 0.00 | 0.00 | 0.50 | 0.49 | 0.26 | 0.00 | 0.00 |
| FFFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.03 | 0.00 | 0.00 | 0.00 | 0.70 | 0.53 | 0.26 | 0.00 | 0.00 | 0.50 | 0.49 | 0.26 | 0.00 | 0.00 |
| FFFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.03 | 0.00 | 0.00 | 0.00 | 0.70 | 0.53 | 0.26 | 0.00 | 0.00 | 0.50 | 0.49 | 0.26 | 0.00 | 0.00 |
| FFFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.03 | 0.00 | 0.00 | 0.00 | 0.70 | 0.53 | 0.26 | 0.00 | 0.00 | 0.50 | 0.49 | 0.26 | 0.00 | 0.00 |

Table A1.12: Causal Responsibility for voters in simple variant of Engl Model - Reward Treatment (Allocation 9,1 vs. 5,5)

| | Ex-post Responsibility (U) | | | | | Ex-ante Responsibility (U) | | | | | Ex-post Responsibility (F) | | | | | Ex-ante Responsibility (F) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | V1 | V2 | V3 | V4 | V5 | V1 | V2 | V3 | V4 | V5 | V1 | V2 | V3 | V4 | V5 | V1 | V2 | V3 | V4 | V5 |
| UUUUU | 0.61 | 0.44 | 0.20 | 0.00 | 0.00 | 0.42 | 0.40 | 0.20 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.12 | 0.06 | 0.00 | 0.00 | 0.00 |
| UUUUF | 0.61 | 0.44 | 0.20 | 0.00 | 0.00 | 0.42 | 0.40 | 0.20 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.12 | 0.06 | 0.00 | 0.00 | 0.00 |
| UUUFU | 0.61 | 0.44 | 0.20 | 0.00 | 0.00 | 0.42 | 0.40 | 0.20 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.12 | 0.06 | 0.00 | 0.00 | 0.00 |
| UUUFF | 0.61 | 0.44 | 0.20 | 0.00 | 0.00 | 0.42 | 0.40 | 0.20 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.12 | 0.06 | 0.00 | 0.00 | 0.00 |
| UUFUU | 0.61 | 0.44 | 0.00 | 0.40 | 0.00 | 0.42 | 0.40 | 0.00 | 0.40 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.12 | 0.06 | 0.20 | 0.00 | 0.00 |
| UUFUF | 0.61 | 0.44 | 0.00 | 0.40 | 0.00 | 0.42 | 0.40 | 0.00 | 0.40 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.12 | 0.06 | 0.20 | 0.00 | 0.00 |
| UUFFU | 0.61 | 0.44 | 0.00 | 0.00 | 1.00 | 0.42 | 0.40 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.12 | 0.06 | 0.20 | 0.40 | 0.00 |
| UUFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.42 | 0.40 | 0.00 | 0.00 | 0.00 | 0.39 | 0.56 | 1.00 | 1.00 | 1.00 | 0.12 | 0.06 | 0.20 | 0.40 | 1.00 |
| UFUUU | 0.61 | 0.10 | 0.67 | 0.40 | 0.00 | 0.42 | 0.06 | 0.56 | 0.40 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.12 | 0.40 | 0.06 | 0.00 | 0.00 |
| UFUUF | 0.61 | 0.10 | 0.67 | 0.40 | 0.00 | 0.42 | 0.06 | 0.56 | 0.40 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.12 | 0.40 | 0.06 | 0.00 | 0.00 |
| UFUFU | 0.61 | 0.10 | 0.67 | 0.00 | 1.00 | 0.42 | 0.06 | 0.56 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.12 | 0.40 | 0.06 | 0.40 | 0.00 |
| UFUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.42 | 0.06 | 0.56 | 0.00 | 0.00 | 0.39 | 0.90 | 0.33 | 1.00 | 1.00 | 0.12 | 0.40 | 0.06 | 0.40 | 1.00 |
| UFFUU | 0.61 | 0.10 | 0.17 | 1.00 | 1.00 | 0.42 | 0.06 | 0.06 | 0.70 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.12 | 0.40 | 0.56 | 0.00 | 0.00 |
| UFFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.42 | 0.06 | 0.06 | 0.70 | 0.00 | 0.39 | 0.90 | 0.83 | 0.00 | 1.00 | 0.12 | 0.40 | 0.56 | 0.00 | 1.00 |
| UFFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.42 | 0.06 | 0.06 | 0.00 | 0.00 | 0.39 | 0.90 | 0.83 | 0.70 | 0.00 | 0.12 | 0.40 | 0.56 | 0.70 | 0.00 |
| UFFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.42 | 0.06 | 0.06 | 0.00 | 0.00 | 0.39 | 0.90 | 0.83 | 0.70 | 0.00 | 0.12 | 0.40 | 0.56 | 0.70 | 0.00 |
| FUUUU | 0.32 | 0.85 | 0.58 | 0.30 | 0.00 | 0.10 | 0.55 | 0.50 | 0.30 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.42 | 0.05 | 0.06 | 0.00 | 0.00 |
| FUUUF | 0.32 | 0.85 | 0.58 | 0.30 | 0.00 | 0.10 | 0.55 | 0.50 | 0.30 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.42 | 0.05 | 0.06 | 0.00 | 0.00 |
| FUUFU | 0.32 | 0.85 | 0.58 | 0.00 | 1.00 | 0.10 | 0.55 | 0.50 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.42 | 0.05 | 0.06 | 0.30 | 0.00 |
| FUUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.55 | 0.50 | 0.00 | 0.00 | 0.68 | 0.15 | 0.42 | 1.00 | 1.00 | 0.42 | 0.05 | 0.06 | 0.30 | 1.00 |
| FUFUU | 0.32 | 0.85 | 0.14 | 1.00 | 1.00 | 0.10 | 0.55 | 0.06 | 0.73 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.42 | 0.05 | 0.50 | 0.00 | 0.00 |
| FUFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.55 | 0.06 | 0.73 | 0.00 | 0.68 | 0.15 | 0.86 | 0.00 | 1.00 | 0.42 | 0.05 | 0.50 | 0.00 | 1.00 |
| FUFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.55 | 0.06 | 0.00 | 0.00 | 0.68 | 0.15 | 0.86 | 0.73 | 0.00 | 0.42 | 0.05 | 0.50 | 0.73 | 0.00 |
| FUFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.55 | 0.06 | 0.00 | 0.00 | 0.68 | 0.15 | 0.86 | 0.73 | 0.00 | 0.42 | 0.05 | 0.50 | 0.73 | 0.00 |
| FFUUU | 0.32 | 0.35 | 1.00 | 1.00 | 1.00 | 0.10 | 0.05 | 0.36 | 0.67 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.42 | 0.55 | 0.00 | 0.00 | 0.00 |
| FFUUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.05 | 0.36 | 0.67 | 0.00 | 0.68 | 0.65 | 0.00 | 0.00 | 1.00 | 0.42 | 0.55 | 0.00 | 0.00 | 1.00 |
| FFUFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.05 | 0.36 | 0.00 | 0.00 | 0.68 | 0.65 | 0.00 | 0.00 | 0.00 | 0.42 | 0.55 | 0.00 | 0.60 | 0.00 |
| FFUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.05 | 0.36 | 0.00 | 0.00 | 0.68 | 0.65 | 0.00 | 0.67 | 0.00 | 0.42 | 0.55 | 0.00 | 0.60 | 0.00 |
| FFFUU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.05 | 0.00 | 0.00 | 0.00 | 0.68 | 0.65 | 0.36 | 0.00 | 0.00 | 0.42 | 0.55 | 0.36 | 0.00 | 0.00 |
| FFFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.05 | 0.00 | 0.00 | 0.00 | 0.68 | 0.65 | 0.36 | 0.00 | 0.00 | 0.42 | 0.55 | 0.36 | 0.00 | 0.00 |
| FFFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.05 | 0.00 | 0.00 | 0.00 | 0.68 | 0.65 | 0.36 | 0.00 | 0.00 | 0.42 | 0.55 | 0.36 | 0.00 | 0.00 |
| FFFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.05 | 0.00 | 0.00 | 0.00 | 0.68 | 0.65 | 0.36 | 0.00 | 0.00 | 0.42 | 0.55 | 0.36 | 0.00 | 0.00 |

Table A1.13: Causal Responsibility for voters in simple variant of Engl Model - Reward Treatment (Allocation 8,2 vs. 6,4)

| | Ex-post Responsibility (U) | | | | | Ex-ante Responsibility (U) | | | | | Ex-post Responsibility (F) | | | | | Ex-ante Responsibility (F) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | V1 | V2 | V3 | V4 | V5 | V1 | V2 | V3 | V4 | V5 | V1 | V2 | V3 | V4 | V5 | V1 | V2 | V3 | V4 | V5 |
| UUUUU | 0.72 | 0.54 | 0.20 | 0.00 | 0.00 | 0.47 | 0.48 | 0.20 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.05 | 0.00 | 0.00 | 0.00 |
| UUUUF | 0.72 | 0.54 | 0.20 | 0.00 | 0.00 | 0.47 | 0.48 | 0.20 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.05 | 0.00 | 0.00 | 0.00 |
| UUUFU | 0.72 | 0.54 | 0.20 | 0.00 | 0.00 | 0.47 | 0.48 | 0.20 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.05 | 0.00 | 0.00 | 0.00 |
| UUUFF | 0.72 | 0.54 | 0.20 | 0.00 | 0.00 | 0.47 | 0.48 | 0.20 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.05 | 0.00 | 0.00 | 0.00 |
| UUFUU | 0.72 | 0.54 | 0.00 | 0.47 | 0.00 | 0.47 | 0.48 | 0.00 | 0.47 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.05 | 0.20 | 0.00 | 0.00 |
| UUFUF | 0.72 | 0.54 | 0.00 | 0.47 | 0.00 | 0.47 | 0.48 | 0.00 | 0.47 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.05 | 0.20 | 0.00 | 0.00 |
| UUFFU | 0.72 | 0.54 | 0.00 | 0.00 | 1.00 | 0.47 | 0.48 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.05 | 0.20 | 0.47 | 0.00 |
| UUFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.47 | 0.48 | 0.00 | 0.00 | 0.00 | 0.28 | 0.46 | 1.00 | 1.00 | 1.00 | 0.10 | 0.05 | 0.20 | 0.47 | 1.00 |
| UFUUU | 0.72 | 0.11 | 0.72 | 0.57 | 0.00 | 0.47 | 0.05 | 0.47 | 0.57 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.48 | 0.10 | 0.00 | 0.00 |
| UFUUF | 0.72 | 0.11 | 0.72 | 0.57 | 0.00 | 0.47 | 0.05 | 0.47 | 0.57 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.48 | 0.10 | 0.00 | 0.00 |
| UFUFU | 0.72 | 0.11 | 0.72 | 0.00 | 1.00 | 0.47 | 0.05 | 0.47 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.48 | 0.10 | 0.57 | 0.00 |
| UFUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.47 | 0.05 | 0.47 | 0.00 | 0.00 | 0.28 | 0.89 | 0.28 | 1.00 | 1.00 | 0.10 | 0.48 | 0.10 | 0.57 | 1.00 |
| UFFUU | 0.72 | 0.11 | 0.34 | 1.00 | 1.00 | 0.47 | 0.05 | 0.10 | 0.57 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.48 | 0.47 | 0.00 | 0.00 |
| UFFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.47 | 0.05 | 0.10 | 0.57 | 0.00 | 0.28 | 0.89 | 0.66 | 0.00 | 1.00 | 0.10 | 0.48 | 0.47 | 0.00 | 1.00 |
| UFFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.47 | 0.05 | 0.10 | 0.00 | 0.00 | 0.28 | 0.89 | 0.66 | 0.57 | 0.00 | 0.10 | 0.48 | 0.47 | 0.57 | 0.00 |
| UFFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.47 | 0.05 | 0.10 | 0.00 | 0.00 | 0.28 | 0.89 | 0.66 | 0.57 | 0.00 | 0.10 | 0.48 | 0.47 | 0.57 | 0.00 |
| FUUUU | 0.35 | 0.89 | 0.72 | 0.40 | 0.00 | 0.08 | 0.51 | 0.61 | 0.40 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.47 | 0.05 | 0.04 | 0.00 | 0.00 |
| FUUUF | 0.35 | 0.89 | 0.72 | 0.40 | 0.00 | 0.08 | 0.51 | 0.61 | 0.40 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.47 | 0.05 | 0.04 | 0.00 | 0.00 |
| FUUFU | 0.35 | 0.89 | 0.72 | 0.00 | 1.00 | 0.08 | 0.51 | 0.61 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.47 | 0.05 | 0.04 | 0.40 | 0.00 |
| FUUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.51 | 0.61 | 0.00 | 0.00 | 0.65 | 0.11 | 0.28 | 1.00 | 1.00 | 0.47 | 0.05 | 0.04 | 0.40 | 1.00 |
| FUFUU | 0.35 | 0.89 | 0.15 | 1.00 | 1.00 | 0.08 | 0.51 | 0.04 | 0.57 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.47 | 0.05 | 0.61 | 0.00 | 0.00 |
| FUFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.51 | 0.04 | 0.57 | 0.00 | 0.65 | 0.11 | 0.85 | 0.00 | 1.00 | 0.47 | 0.05 | 0.61 | 0.00 | 1.00 |
| FUFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.51 | 0.04 | 0.00 | 0.00 | 0.65 | 0.11 | 0.85 | 0.57 | 0.00 | 0.47 | 0.05 | 0.61 | 0.57 | 0.00 |
| FUFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.51 | 0.04 | 0.00 | 0.00 | 0.65 | 0.11 | 0.85 | 0.57 | 0.00 | 0.47 | 0.05 | 0.61 | 0.57 | 0.00 |
| FFUUU | 0.35 | 0.43 | 1.00 | 1.00 | 1.00 | 0.08 | 0.05 | 0.30 | 0.57 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.47 | 0.51 | 0.00 | 0.00 | 0.00 |
| FFUUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.05 | 0.30 | 0.57 | 0.00 | 0.65 | 0.57 | 0.00 | 0.00 | 1.00 | 0.47 | 0.51 | 0.00 | 0.00 | 1.00 |
| FFUFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.05 | 0.30 | 0.00 | 0.00 | 0.65 | 0.57 | 0.00 | 0.00 | 0.00 | 0.47 | 0.51 | 0.00 | 0.49 | 0.00 |
| FFUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.05 | 0.30 | 0.00 | 0.00 | 0.65 | 0.57 | 0.00 | 0.57 | 0.00 | 0.47 | 0.51 | 0.00 | 0.49 | 0.00 |
| FFFUU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.05 | 0.00 | 0.00 | 0.00 | 0.65 | 0.57 | 0.30 | 0.00 | 0.00 | 0.47 | 0.51 | 0.30 | 0.00 | 0.00 |
| FFFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.05 | 0.00 | 0.00 | 0.00 | 0.65 | 0.57 | 0.30 | 0.00 | 0.00 | 0.47 | 0.51 | 0.30 | 0.00 | 0.00 |
| FFFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.05 | 0.00 | 0.00 | 0.00 | 0.65 | 0.57 | 0.30 | 0.00 | 0.00 | 0.47 | 0.51 | 0.30 | 0.00 | 0.00 |
| FFFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.05 | 0.00 | 0.00 | 0.00 | 0.65 | 0.57 | 0.30 | 0.00 | 0.00 | 0.47 | 0.51 | 0.30 | 0.00 | 0.00 |

Table A1.14: Causal Responsibility for voters in simple variant of Engl Model - Both Treatment (Allocation 9,1 vs. 5,5)

| | Ex-post Responsibility (U) | | | | | Ex-ante Responsibility (U) | | | | | Ex-post Responsibility (F) | | | | | Ex-ante Responsibility (F) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | V1 | V2 | V3 | V4 | V5 | V1 | V2 | V3 | V4 | V5 | V1 | V2 | V3 | V4 | V5 | V1 | V2 | V3 | V4 | V5 |
| UUUUU | 0.72 | 0.57 | 0.30 | 0.00 | 0.00 | 0.45 | 0.47 | 0.30 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.11 | 0.08 | 0.00 | 0.00 | 0.00 |
| UUUUF | 0.72 | 0.57 | 0.30 | 0.00 | 0.00 | 0.45 | 0.47 | 0.30 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.11 | 0.08 | 0.00 | 0.00 | 0.00 |
| UUUFU | 0.72 | 0.57 | 0.30 | 0.00 | 0.00 | 0.45 | 0.47 | 0.30 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.11 | 0.08 | 0.00 | 0.00 | 0.00 |
| UUUFF | 0.72 | 0.57 | 0.30 | 0.00 | 0.00 | 0.45 | 0.47 | 0.30 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.11 | 0.08 | 0.00 | 0.00 | 0.00 |
| UUFUU | 0.72 | 0.57 | 0.00 | 0.47 | 0.00 | 0.45 | 0.47 | 0.00 | 0.47 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.11 | 0.08 | 0.30 | 0.00 | 0.00 |
| UUFUF | 0.72 | 0.57 | 0.00 | 0.47 | 0.00 | 0.45 | 0.47 | 0.00 | 0.47 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.11 | 0.08 | 0.30 | 0.00 | 0.00 |
| UUFFU | 0.72 | 0.57 | 0.00 | 0.00 | 1.00 | 0.45 | 0.47 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.11 | 0.08 | 0.30 | 0.47 | 0.00 |
| UUFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.45 | 0.47 | 0.00 | 0.00 | 0.00 | 0.28 | 0.43 | 1.00 | 1.00 | 1.00 | 0.11 | 0.08 | 0.30 | 0.47 | 1.00 |
| UFUUU | 0.72 | 0.18 | 0.81 | 0.47 | 0.00 | 0.45 | 0.08 | 0.61 | 0.47 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.11 | 0.47 | 0.05 | 0.00 | 0.00 |
| UFUUF | 0.72 | 0.18 | 0.81 | 0.47 | 0.00 | 0.45 | 0.08 | 0.61 | 0.47 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.11 | 0.47 | 0.05 | 0.00 | 0.00 |
| UFUFU | 0.72 | 0.18 | 0.81 | 0.00 | 1.00 | 0.45 | 0.08 | 0.61 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.11 | 0.47 | 0.05 | 0.47 | 0.00 |
| UFUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.45 | 0.08 | 0.61 | 0.00 | 0.00 | 0.28 | 0.82 | 0.19 | 1.00 | 1.00 | 0.11 | 0.47 | 0.05 | 0.47 | 1.00 |
| UFFUU | 0.72 | 0.18 | 0.25 | 1.00 | 1.00 | 0.45 | 0.08 | 0.05 | 0.43 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.11 | 0.47 | 0.61 | 0.00 | 0.00 |
| UFFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.45 | 0.08 | 0.05 | 0.43 | 0.00 | 0.28 | 0.82 | 0.75 | 0.00 | 1.00 | 0.11 | 0.47 | 0.61 | 0.00 | 1.00 |
| UFFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.45 | 0.08 | 0.05 | 0.00 | 0.00 | 0.28 | 0.82 | 0.75 | 0.43 | 0.00 | 0.11 | 0.47 | 0.61 | 0.43 | 0.00 |
| UFFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.45 | 0.08 | 0.05 | 0.00 | 0.00 | 0.28 | 0.82 | 0.75 | 0.43 | 0.00 | 0.11 | 0.47 | 0.61 | 0.43 | 0.00 |
| FUUUU | 0.37 | 0.93 | 0.72 | 0.43 | 0.00 | 0.07 | 0.49 | 0.53 | 0.43 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.45 | 0.03 | 0.07 | 0.00 | 0.00 |
| FUUUF | 0.37 | 0.93 | 0.72 | 0.43 | 0.00 | 0.07 | 0.49 | 0.53 | 0.43 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.45 | 0.03 | 0.07 | 0.00 | 0.00 |
| FUUFU | 0.37 | 0.93 | 0.72 | 0.00 | 1.00 | 0.07 | 0.49 | 0.53 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.45 | 0.03 | 0.07 | 0.43 | 0.00 |
| FUUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.07 | 0.49 | 0.53 | 0.00 | 0.00 | 0.63 | 0.07 | 0.28 | 1.00 | 1.00 | 0.45 | 0.03 | 0.07 | 0.43 | 1.00 |
| FUFUU | 0.37 | 0.93 | 0.26 | 1.00 | 1.00 | 0.07 | 0.49 | 0.07 | 0.50 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.45 | 0.03 | 0.53 | 0.00 | 0.00 |
| FUFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.07 | 0.49 | 0.07 | 0.50 | 0.00 | 0.63 | 0.07 | 0.74 | 0.00 | 1.00 | 0.45 | 0.03 | 0.53 | 0.00 | 1.00 |
| FUFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.07 | 0.49 | 0.07 | 0.00 | 0.00 | 0.63 | 0.07 | 0.74 | 0.50 | 0.00 | 0.45 | 0.03 | 0.53 | 0.50 | 0.00 |
| FUFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.07 | 0.49 | 0.07 | 0.00 | 0.00 | 0.63 | 0.07 | 0.74 | 0.50 | 0.00 | 0.45 | 0.03 | 0.53 | 0.50 | 0.00 |
| FFUUU | 0.37 | 0.47 | 1.00 | 1.00 | 1.00 | 0.07 | 0.03 | 0.21 | 0.53 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.45 | 0.49 | 0.00 | 0.00 | 0.00 |
| FFUUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.07 | 0.03 | 0.21 | 0.53 | 0.00 | 0.63 | 0.53 | 0.00 | 0.00 | 1.00 | 0.45 | 0.49 | 0.00 | 0.00 | 1.00 |
| FFUFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.07 | 0.03 | 0.21 | 0.00 | 0.00 | 0.63 | 0.53 | 0.00 | 0.00 | 0.00 | 0.45 | 0.49 | 0.00 | 0.52 | 0.00 |
| FFUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.07 | 0.03 | 0.21 | 0.00 | 0.00 | 0.63 | 0.53 | 0.00 | 0.53 | 0.00 | 0.45 | 0.49 | 0.00 | 0.52 | 0.00 |
| FFFUU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.07 | 0.03 | 0.00 | 0.00 | 0.00 | 0.63 | 0.53 | 0.21 | 0.00 | 0.00 | 0.45 | 0.49 | 0.21 | 0.00 | 0.00 |
| FFFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.07 | 0.03 | 0.00 | 0.00 | 0.00 | 0.63 | 0.53 | 0.21 | 0.00 | 0.00 | 0.45 | 0.49 | 0.21 | 0.00 | 0.00 |
| FFFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.07 | 0.03 | 0.00 | 0.00 | 0.00 | 0.63 | 0.53 | 0.21 | 0.00 | 0.00 | 0.45 | 0.49 | 0.21 | 0.00 | 0.00 |
| FFFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.07 | 0.03 | 0.00 | 0.00 | 0.00 | 0.63 | 0.53 | 0.21 | 0.00 | 0.00 | 0.45 | 0.49 | 0.21 | 0.00 | 0.00 |

Table A1.15: Causal Responsibility for voters in simple variant of Engl Model - Both Treatment (Allocation 8,2 vs. 6,4)

| | Ex-post Responsibility (U) | | | | | Ex-ante Responsibility (U) | | | | | Ex-post Responsibility (F) | | | | | Ex-ante Responsibility (F) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | V1 | V2 | V3 | V4 | V5 | V1 | V2 | V3 | V4 | V5 | V1 | V2 | V3 | V4 | V5 | V1 | V2 | V3 | V4 | V5 |
| UUUUU | 0.74 | 0.58 | 0.40 | 0.00 | 0.00 | 0.46 | 0.44 | 0.40 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.10 | 0.00 | 0.00 | 0.00 |
| UUUUF | 0.74 | 0.58 | 0.40 | 0.00 | 0.00 | 0.46 | 0.44 | 0.40 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.10 | 0.00 | 0.00 | 0.00 |
| UUUFU | 0.74 | 0.58 | 0.40 | 0.00 | 0.00 | 0.46 | 0.44 | 0.40 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.10 | 0.00 | 0.00 | 0.00 |
| UUUFF | 0.74 | 0.58 | 0.40 | 0.00 | 0.00 | 0.46 | 0.44 | 0.40 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.10 | 0.00 | 0.00 | 0.00 |
| UUFUU | 0.74 | 0.58 | 0.00 | 0.63 | 0.00 | 0.46 | 0.44 | 0.00 | 0.63 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.10 | 0.40 | 0.00 | 0.00 |
| UUFUF | 0.74 | 0.58 | 0.00 | 0.63 | 0.00 | 0.46 | 0.44 | 0.00 | 0.63 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.10 | 0.40 | 0.00 | 0.00 |
| UUFFU | 0.74 | 0.58 | 0.00 | 0.00 | 1.00 | 0.46 | 0.44 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.10 | 0.40 | 0.63 | 0.00 |
| UUFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.46 | 0.44 | 0.00 | 0.00 | 0.00 | 0.26 | 0.42 | 1.00 | 1.00 | 1.00 | 0.10 | 0.10 | 0.40 | 0.63 | 1.00 |
| UFUUU | 0.74 | 0.24 | 0.80 | 0.57 | 0.00 | 0.46 | 0.10 | 0.54 | 0.57 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.44 | 0.06 | 0.00 | 0.00 |
| UFUUF | 0.74 | 0.24 | 0.80 | 0.57 | 0.00 | 0.46 | 0.10 | 0.54 | 0.57 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.44 | 0.06 | 0.00 | 0.00 |
| UFUFU | 0.74 | 0.24 | 0.80 | 0.00 | 1.00 | 0.46 | 0.10 | 0.54 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.44 | 0.06 | 0.57 | 0.00 |
| UFUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.46 | 0.10 | 0.54 | 0.00 | 0.00 | 0.26 | 0.76 | 0.20 | 1.00 | 1.00 | 0.10 | 0.44 | 0.06 | 0.57 | 1.00 |
| UFFUU | 0.74 | 0.24 | 0.32 | 1.00 | 1.00 | 0.46 | 0.10 | 0.06 | 0.43 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.44 | 0.54 | 0.00 | 0.00 |
| UFFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.46 | 0.10 | 0.06 | 0.43 | 0.00 | 0.26 | 0.76 | 0.68 | 0.00 | 1.00 | 0.10 | 0.44 | 0.54 | 0.00 | 1.00 |
| UFFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.46 | 0.10 | 0.06 | 0.00 | 0.00 | 0.26 | 0.76 | 0.68 | 0.43 | 0.00 | 0.10 | 0.44 | 0.54 | 0.43 | 0.00 |
| UFFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.46 | 0.10 | 0.06 | 0.00 | 0.00 | 0.26 | 0.76 | 0.68 | 0.43 | 0.00 | 0.10 | 0.44 | 0.54 | 0.43 | 0.00 |
| FUUUU | 0.37 | 0.92 | 0.80 | 0.53 | 0.00 | 0.08 | 0.41 | 0.57 | 0.53 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.46 | 0.04 | 0.06 | 0.00 | 0.00 |
| FUUUF | 0.37 | 0.92 | 0.80 | 0.53 | 0.00 | 0.08 | 0.41 | 0.57 | 0.53 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.46 | 0.04 | 0.06 | 0.00 | 0.00 |
| FUUFU | 0.37 | 0.92 | 0.80 | 0.00 | 1.00 | 0.08 | 0.41 | 0.57 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.46 | 0.04 | 0.06 | 0.53 | 0.00 |
| FUUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.41 | 0.57 | 0.00 | 0.00 | 0.63 | 0.08 | 0.20 | 1.00 | 1.00 | 0.46 | 0.04 | 0.06 | 0.53 | 1.00 |
| FUFUU | 0.37 | 0.92 | 0.28 | 1.00 | 1.00 | 0.08 | 0.41 | 0.06 | 0.40 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.46 | 0.04 | 0.57 | 0.00 | 0.00 |
| FUFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.41 | 0.06 | 0.40 | 0.00 | 0.63 | 0.08 | 0.72 | 0.00 | 1.00 | 0.46 | 0.04 | 0.57 | 0.00 | 1.00 |
| FUFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.41 | 0.06 | 0.00 | 0.00 | 0.63 | 0.08 | 0.72 | 0.40 | 0.00 | 0.46 | 0.04 | 0.57 | 0.40 | 0.00 |
| FUFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.41 | 0.06 | 0.00 | 0.00 | 0.63 | 0.08 | 0.72 | 0.40 | 0.00 | 0.46 | 0.04 | 0.57 | 0.40 | 0.00 |
| FFUUU | 0.37 | 0.56 | 1.00 | 1.00 | 1.00 | 0.08 | 0.04 | 0.22 | 0.50 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.46 | 0.41 | 0.00 | 0.00 | 0.00 |
| FFUUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.04 | 0.22 | 0.50 | 0.00 | 0.63 | 0.44 | 0.00 | 0.00 | 1.00 | 0.46 | 0.41 | 0.00 | 0.00 | 1.00 |
| FFUFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.04 | 0.22 | 0.00 | 0.00 | 0.63 | 0.44 | 0.00 | 0.00 | 0.00 | 0.46 | 0.41 | 0.00 | 0.48 | 0.00 |
| FFUFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.04 | 0.22 | 0.00 | 0.00 | 0.63 | 0.44 | 0.00 | 0.50 | 0.00 | 0.46 | 0.41 | 0.00 | 0.48 | 0.00 |
| FFFUU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.04 | 0.00 | 0.00 | 0.00 | 0.63 | 0.44 | 0.22 | 0.00 | 0.00 | 0.46 | 0.41 | 0.22 | 0.00 | 0.00 |
| FFFUF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.04 | 0.00 | 0.00 | 0.00 | 0.63 | 0.44 | 0.22 | 0.00 | 0.00 | 0.46 | 0.41 | 0.22 | 0.00 | 0.00 |
| FFFFU | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.04 | 0.00 | 0.00 | 0.00 | 0.63 | 0.44 | 0.22 | 0.00 | 0.00 | 0.46 | 0.41 | 0.22 | 0.00 | 0.00 |
| FFFFF | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.04 | 0.00 | 0.00 | 0.00 | 0.63 | 0.44 | 0.22 | 0.00 | 0.00 | 0.46 | 0.41 | 0.22 | 0.00 | 0.00 |

### 1.6.4 Correlation Matrix - Sanction Motives

Table A1.16: Punishment Treatment

| Punishment | Choice Unfair | Outcome Unfair | Intention Unkind | Initiator Unfair | Pivotal Unfair | BF Responsibility(U) | Ex-ante Engl Resp(U) | Ex-post Engl Resp(U) |
|---|---|---|---|---|---|---|---|---|
| Choice Unfair | 1 | | | | | | | |
| Outcome Unfair | 0.375 | 1 | | | | | | |
| Intention Unkind | 0.838 | 0.381 | 1 | | | | | |
| Initiator Unfair | 0.333 | 0.333 | 0.398 | 1 | | | | |
| Pivotal Unfair | 0.333 | 0.333 | 0.398 | -0.111 | 1 | | | |
| BF Responsibility(U) | 0.616 | 0.616 | 0.735 | 0.502 | 0.549 | 1 | | |
| Ex-ante Engl Resp(U) | 0.739 | 0.369 | 0.897 | 0.337 | 0.439 | 0.789 | 1 | |
| Ex-post Engl Resp(U) | 0.561 | 0.653 | 0.686 | 0.538 | 0.367 | 0.905 | 0.773 | 1 |

Table A1.17: Reward Treatment

| Reward | Choice Fair | Outcome Fair | Intention Kind | Initiator Fair | Pivotal Fair | BF Responsibility(F) | Ex-ante Engl Resp(F) | Ex-post Engl Resp(F) |
|---|---|---|---|---|---|---|---|---|
| Choice Fair | 1 | | | | | | | |
| Outcome Fair | 0.375 | 1 | | | | | | |
| Intention Kind | 0.838 | 0.381 | 1 | | | | | |
| Initiator Fair | 0.333 | 0.333 | 0.398 | 1 | | | | |
| Pivotal Fair | 0.333 | 0.333 | 0.398 | -0.111 | 1 | | | |
| BF Responsibility(F) | 0.576 | 0.576 | 0.687 | 0.210 | 0.729 | 1 | | |
| Ex-ante Engl Resp(F) | 0.717 | 0.409 | 0.882 | 0.251 | 0.577 | 0.850 | 1 | |
| Ex-post Engl Resp(F) | 0.561 | 0.682 | 0.691 | 0.488 | 0.403 | 0.824 | 0.765 | 1 |

Table A1.18: Both Treatment - Unfair Variables

| Both | Choice Unfair | Outcome Unfair | Intention Unkind | Initiator Unfair | Pivotal Unfair | BF Responsibility(U) | Ex-ante Engl Resp(U) | Ex-post Engl Resp(U) |
|---|---|---|---|---|---|---|---|---|
| Choice Unfair | 1 | | | | | | | |
| Outcome Unfair | 0.375 | 1 | | | | | | |
| Intention Unkind | 0.838 | 0.381 | 1 | | | | | |
| Initiator Unfair | 0.333 | 0.333 | 0.398 | 1 | | | | |
| Pivotal Unfair | 0.333 | 0.333 | 0.398 | -0.111 | 1 | | | |
| BF Responsibility(U) | 0.605 | 0.605 | 0.722 | 0.329 | 0.711 | 1 | | |
| Ex-ante Resp(U) | 0.728 | 0.403 | 0.892 | 0.274 | 0.551 | 0.828 | 1 | |
| Ex-post Resp(U) | 0.560 | 0.696 | 0.692 | 0.530 | 0.397 | 0.862 | 0.765 | 1 |

Table A1.19: Both Treatment - Fair Variables

| Both | Choice Fair | Outcome Fair | Intention Kind | Initiator Fair | Pivotal Fair | BF Responsibility(F) | Ex-ante Engl Resp(F) | Ex-post Engl Resp(F) |
|---|---|---|---|---|---|---|---|---|
| Choice Fair | 1 | | | | | | | |
| Outcome Fair | 0.375 | 1 | | | | | | |
| Intention Kind | 0.838 | 0.381 | 1 | | | | | |
| Initiator Fair | 0.333 | 0.333 | 0.398 | 1 | | | | |
| Pivotal Fair | 0.333 | 0.333 | 0.398 | -0.111 | 1 | | | |
| BF Responsibility(F) | 0.601 | 0.601 | 0.717 | 0.344 | 0.565 | 1 | | |
| Ex-ante Engl Resp(F) | 0.722 | 0.366 | 0.888 | 0.285 | 0.481 | 0.804 | 1 | |
| Ex-post Engl Resp(F) | 0.552 | 0.650 | 0.676 | 0.500 | 0.376 | 0.869 | 0.772 | 1 |

## 1.6.5 Individual Regressions - Theoretical Framework

Table A1.20: Punishment Treatment

| Pun. Points | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 | Model 7 | Model 8 | Model 9 | Model 10 | Model 11 | Model 12 | Model 13 | Model 14 | Model 15 | Model 16 | Model 17 | Model 18 | Model 19 | Model 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Choice Unfair | 1.013*** | | | | | | | | | | | | | 0.770*** | | | | 0.759*** | 0.799*** | 0.772*** |
| | (0.123) | | | | | | | | | | | | | (0.177) | | | | (0.161) | (0.176) | (0.154) |
| Outcome Unfair | | 0.529*** | | | | | | | | | | | | 0.0336 | | | | -0.115 | -0.0902 | -0.112 |
| | | (0.111) | | | | | | | | | | | | (0.099) | | | | (0.107) | (0.061) | (0.110) |
| Intention Unkind | | | 0.947*** | | | | | | | | | | | 0.051 | | | | 0.075 | 0.073 | -0.008 |
| | | | (0.124) | | | | | | | | | | | (0.130) | | | | (0.151) | (0.152) | (0.135) |
| Initiator Unfair | | | | 0.705** | | | | | | | | | | 0.296 | | | | 0.146 | 0.145 | 0.183 |
| | | | | (0.244) | | | | | | | | | | (0.233) | | | | (0.230) | (0.230) | (0.228) |
| Pivotal Unfair | | | | | 1.090*** | | | | | | | | | 0.643** | | | | 0.603** | 0.603** | 0.565** |
| | | | | | (0.240) | | | | | | | | | (0.211) | | | | (0.209) | (0.208) | (0.204) |
| BF Resp (U) | | | | | | 2.698*** | | | | | | | | | 1.950** | 1.594** | 1.327* | 0.084 | 0.067 | 0.464 |
| | | | | | | (0.431) | | | | | | | | | (0.538) | (0.553) | (0.523) | (0.358) | (0.416) | (0.446) |
| EA Engl (U) | | | | | | | 1.561*** | | | | | | | | 0.871* | 0.525 | | -0.384 | -0.374 | |
| | | | | | | | (0.214) | | | | | | | | (0.329) | (0.297) | | (0.208) | (0.215) | |
| EA Engl (F) | | | | | | | | -1.380*** | | | | | | | | -0.739*** | | -0.075 | | |
| | | | | | | | | (0.168) | | | | | | | | (0.108) | | (0.053) | | |
| EA Engl (U-F) | | | | | | | | | 0.971*** | | | | | | | | 0.655** | | | -0.109 |
| | | | | | | | | | (0.123) | | | | | | | | (0.189) | | | (0.055) |
| EP Engl (U) | | | | | | | | | | 1.175*** | | | | | -0.225 | -0.080 | | 0.484** | 0.467** | |
| | | | | | | | | | | (0.194) | | | | | (0.237) | (0.241) | | (0.156) | (0.152) | |
| EP Engl (F) | | | | | | | | | | | -0.867*** | | | | | 0.008 | | -0.029 | | |
| | | | | | | | | | | | (0.102) | | | | | (0.030) | | (0.112) | | |
| EP Engl (U-F) | | | | | | | | | | | | 0.709*** | | | | | -0.001 | | | 0.175* |
| | | | | | | | | | | | | (0.096) | | | | | (0.071) | | | (0.078) |
| Engl Comb. (U) | | | | | | | | | | | | | 1.513*** | | | | | | | |
| | | | | | | | | | | | | | (0.219) | | | | | | | |
| Constant | 0.038 | 0.279*** | 0.153** | 0.474*** | 0.435*** | 0.274*** | 0.188*** | 0.862*** | 0.546*** | 0.285*** | 0.745*** | 0.552*** | 0.205*** | 0.027 | 0.200*** | 0.451*** | 0.413*** | 0.113 | 0.060 | 0.098 |
| | (0.062) | (0.056) | (0.051) | (0.024) | (0.024) | (0.043) | (0.049) | (0.039) | (0.000) | (0.043) | (0.024) | (0.001) | (0.049) | (0.059) | (0.047) | (0.049) | (0.052) | (0.064) | (0.057) | (0.068) |
| Observations | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 |
| R-squared | 0.181 | 0.050 | 0.154 | 0.032 | 0.076 | 0.136 | 0.132 | 0.102 | 0.154 | 0.111 | 0.065 | 0.120 | 0.136 | 0.207 | 0.151 | 0.171 | 0.171 | 0.211 | 0.211 | 0.210 |
| Number of Subjects | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 |

*Note:* Fixed effects regression with punishment points as dependent variable (ranges from 0 to 7). *BF Responsibility Unfair* represents a voters' share in the probability increase of an unfair outcome. *Ex-ante (EA) and Ex-post (EP) Engl Responsibility (Un)fair* represent a voter's causal responsibility for a (un)fair event. We report within R-squared in the table. Robust standard errors in parentheses clustered at the subject level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A1.21: Reward Treatment

| Reward Points | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 | Model 7 | Model 8 | Model 9 | Model 10 | Model 11 | Model 12 | Model 13 | Model 14 | Model 15 | Model 16 | Model 17 | Model 18 | Model 19 | Model 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Choice Fair | 0.758*** | | | | | | | | | | | | | 0.451** | | | | 0.462*** | 0.481*** | 0.448*** |
| | (0.134) | | | | | | | | | | | | | (0.123) | | | | (0.115) | (0.123) | (0.110) |
| Outcome Fair | | 0.408*** | | | | | | | | | | | | 0.080 | | | | -0.012 | -0.011 | -0.037 |
| | | (0.079) | | | | | | | | | | | | (0.061) | | | | (0.076) | (0.041) | (0.070) |
| Intention Kind | | | 0.750*** | | | | | | | | | | | 0.261** | | | | 0.175* | 0.176* | 0.207* |
| | | | (0.135) | | | | | | | | | | | (0.094) | | | | (0.082) | (0.082) | (0.091) |
| Initiator Fair | | | | 0.459*** | | | | | | | | | | 0.025 | | | | -0.054 | -0.055 | -0.003 |
| | | | | (0.090) | | | | | | | | | | (0.064) | | | | (0.058) | (0.063) | (0.066) |
| Pivotal Fair | | | | | 0.744*** | | | | | | | | | 0.281* | | | | 0.275** | 0.276** | 0.216** |
| | | | | | (0.153) | | | | | | | | | (0.115) | | | | (0.082) | (0.082) | (0.077) |
| BF Resp (F) | | | | | | 1.833*** | | | | | | | | | 0.092 | 0.079 | 0.546 | -0.127 | -0.119 | 0.269 |
| | | | | | | (0.348) | | | | | | | | | (0.412) | (0.411) | (0.385) | (0.312) | (0.357) | (0.352) |
| EA Engl (U) | | | | | | | -1.046*** | | | | | | | | | -0.531*** | | -0.055 | | |
| | | | | | | | (0.185) | | | | | | | | | (0.105) | | (0.074) | | |
| EA Engl (F) | | | | | | | | 1.328*** | | | | | | | 1.016** | 0.762* | | -0.045 | -0.048 | |
| | | | | | | | | (0.244) | | | | | | | (0.344) | (0.304) | | (0.165) | (0.165) | |
| EA Engl (F-U) | | | | | | | | | 0.794*** | | | | | | | | 0.562** | | | -0.068 |
| | | | | | | | | | (0.140) | | | | | | | | (0.173) | | | (0.036) |
| EP Engl (U) | | | | | | | | | | -0.691*** | | | | | | 0.011 | | 0.010 | | |
| | | | | | | | | | | (0.120) | | | | | | (0.068) | | (0.126) | | |
| EP Engl (F) | | | | | | | | | | | 0.890*** | | | | 0.259* | 0.261* | | 0.349** | 0.343** | |
| | | | | | | | | | | | (0.163) | | | | (0.097) | (0.104) | | (0.102) | (0.100) | |
| EP Engl (F-U) | | | | | | | | | | | | 0.550*** | | | | | 0.083 | | | 0.144 |
| | | | | | | | | | | | | (0.093) | | | | | (0.048) | | | (0.071) |
| Engl Comb (F) | | | | | | | | | | | | | 1.209*** | | | | | | | |
| | | | | | | | | | | | | | (0.218) | | | | | | | |
| Constant | 0.144* | 0.318*** | 0.213*** | 0.476*** | 0.448*** | 0.339*** | 0.760*** | 0.219*** | 0.521*** | 0.671*** | 0.312*** | 0.511*** | 0.241*** | 0.119 | 0.220*** | 0.397*** | 0.465*** | 0.154* | 0.134 | 0.180** |
| | (0.067) | (0.040) | (0.056) | (0.009) | (0.015) | (0.035) | (0.042) | (0.056) | (0.000) | (0.026) | (0.039) | (0.002) | (0.051) | (0.065) | (0.057) | (0.045) | (0.039) | (0.064) | (0.068) | (0.058) |
| Observations | 9,280 | 9,280 | 9,280 | 9,280 | 9,280 | 9,280 | 9,280 | 9,280 | 9,280 | 9,280 | 9,280 | 9,280 | 9,280 | 9,280 | 9,280 | 9,280 | 9,280 | 9,280 | 9,280 | 9,280 |
| R-squared | 0.143 | 0.042 | 0.136 | 0.019 | 0.050 | 0.101 | 0.078 | 0.123 | 0.133 | 0.051 | 0.095 | 0.099 | 0.121 | 0.161 | 0.127 | 0.142 | 0.139 | 0.164 | 0.164 | 0.163 |
| Number of Subjects | 29 | 29 | 29 | 29 | 29 | 29 | 29 | 29 | 29 | 29 | 29 | 29 | 29 | 29 | 29 | 29 | 29 | 29 | 29 | 29 |

*Note:* Fixed effects regression with reward points as dependent variable (ranges from 0 to 7). *BF Responsibility Fair* represents a voters' share in the probability increase of a fair outcome. *Ex-ante (EA) and Ex-post (EP) Engl Responsibility (Un)fair* represent a voter's causal responsibility for a (un)fair event. We report within R-squared in the table. Robust standard errors in parentheses clustered at the subject level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A1.22: Both Treatment - Punishment Points

| Pun. Points | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 | Model 7 | Model 8 | Model 9 | Model 10 | Model 11 | Model 12 | Model 13 | Model 14 | Model 15 | Model 16 | Model 17 | Model 18 | Model 19 | Model 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Choice Unfair | 0.729*** | | | | | | | | | | | | | 0.481*** | | | | 0.488*** | 0.519*** | 0.452*** |
| | (0.120) | | | | | | | | | | | | | (0.103) | | | | (0.099) | (0.102) | (0.102) |
| Outcome Unfair | | 0.353*** | | | | | | | | | | | | 0.008 | | | | -0.110* | -0.072* | -0.126* |
| | | (0.074) | | | | | | | | | | | | (0.049) | | | | (0.053) | (0.033) | (0.052) |
| Intention Unkind | | | 0.709*** | | | | | | | | | | | 0.177* | | | | 0.036 | 0.039 | 0.114 |
| | | | (0.124) | | | | | | | | | | | (0.077) | | | | (0.111) | (0.111) | (0.083) |
| Initiator Unfair | | | | 0.714** | | | | | | | | | | 0.344 | | | | 0.283 | 0.277 | 0.298 |
| | | | | (0.246) | | | | | | | | | | (0.240) | | | | (0.242) | (0.243) | (0.242) |
| Pivotal Unfair | | | | | 0.498** | | | | | | | | | 0.149 | | | | 0.183* | 0.191* | 0.109 |
| | | | | | (0.142) | | | | | | | | | (0.117) | | | | (0.085) | (0.087) | (0.090) |
| BF Resp (U) | | | | | | 1.733*** | | | | | | | | | -0.242 | -0.405 | 0.334 | -0.394 | -0.461 | 0.140 |
| | | | | | | (0.322) | | | | | | | | | (0.408) | (0.436) | (0.315) | (0.254) | (0.283) | (0.300) |
| EA Engl (U) | | | | | | | 1.233*** | | | | | | | | 0.945*** | 0.686** | | 0.147 | 0.146 | |
| | | | | | | | (0.208) | | | | | | | | (0.242) | (0.210) | | (0.184) | (0.183) | |
| EA Engl (F) | | | | | | | | -1.010*** | | | | | | | | -0.614*** | | -0.009 | | |
| | | | | | | | | (0.162) | | | | | | | | (0.138) | | (0.052) | | |
| EA Engl (U-F) | | | | | | | | | 0.750*** | | | | | | | | 0.532*** | | | -0.023 |
| | | | | | | | | | (0.122) | | | | | | | | (0.142) | | | (0.060) |
| EP Engl (U) | | | | | | | | | | 0.852*** | | | | | 0.408 | 0.485 | | 0.379*** | 0.380*** | |
| | | | | | | | | | | (0.161) | | | | | (0.226) | (0.254) | | (0.083) | (0.080) | |
| EP Engl (F) | | | | | | | | | | | -0.671*** | | | | | 0.079 | | -0.088 | | |
| | | | | | | | | | | | (0.103) | | | | | (0.063) | | (0.066) | | |
| EP Engl (U-F) | | | | | | | | | | | | 0.530*** | | | | | 0.131 | | | 0.185*** |
| | | | | | | | | | | | | (0.090) | | | | | (0.079) | | | (0.045) |
| Engl Comb (U) | | | | | | | | | | | | | 1.141*** | | | | | | | |
| | | | | | | | | | | | | | (0.200) | | | | | | | |
| Constant | 0.027 | 0.215*** | 0.099 | 0.320*** | 0.342*** | 0.218*** | 0.115* | 0.620*** | 0.393*** | 0.185*** | 0.532*** | 0.374*** | 0.126* | 0.025 | 0.105* | 0.283*** | 0.355*** | 0.076 | 0.027 | 0.120 |
| | (0.060) | (0.037) | (0.051) | (0.025) | (0.014) | (0.032) | (0.047) | (0.037) | (0.000) | (0.039) | (0.022) | (0.003) | (0.047) | (0.058) | (0.048) | (0.036) | (0.032) | (0.055) | (0.056) | (0.060) |
| Observations | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 |
| R-squared | 0.133 | 0.031 | 0.122 | 0.046 | 0.022 | 0.082 | 0.104 | 0.069 | 0.115 | 0.088 | 0.047 | 0.092 | 0.107 | 0.148 | 0.110 | 0.125 | 0.120 | 0.152 | 0.151 | 0.151 |
| Number of Subjects | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 |

*Note:* Fixed effects regression with punishment points as dependent variable (ranges from 0 to 7). *BF Responsibility Unfair* represents a voters' share in the probability increase of an unfair outcome. *Ex-ante (EA) and Ex-post (EP) Engl Responsibility (Un)fair* represent a voter's causal responsibility for a (un)fair event. We report within R-squared in the table. Robust standard errors in parentheses clustered at the subject level. $^{*}$ $p < 0.05$, $^{**}$ $p < 0.01$, $^{***}$ $p < 0.001$

Table A1.23: Both Treatment - Reward Points

| Reward Points | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 | Model 7 | Model 8 | Model 9 | Model 10 | Model 11 | Model 12 | Model 13 | Model 14 | Model 15 | Model 16 | Model 17 | Model 18 | Model 19 | Model 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Choice Fair | 0.309*** | | | | | | | | | | | | | 0.374*** | | | | 0.353*** | 0.378*** | 0.339*** |
| | (0.061) | | | | | | | | | | | | | (0.090) | | | | (0.082) | (0.088) | (0.085) |
| Outcome Fair | | 0.015 | | | | | | | | | | | | -0.112* | | | | -0.165** | -0.090** | -0.169** |
| | | (0.051) | | | | | | | | | | | | (0.044) | | | | (0.049) | (0.026) | (0.048) |
| Intention Kind | | | 0.247*** | | | | | | | | | | | -0.022 | | | | -0.066 | -0.073 | -0.010 |
| | | | (0.055) | | | | | | | | | | | (0.043) | | | | (0.053) | (0.055) | (0.038) |
| Initiator Fair | | | | 0.090 | | | | | | | | | | -0.040 | | | | -0.012 | -0.021 | -0.041 |
| | | | | (0.051) | | | | | | | | | | (0.034) | | | | (0.025) | (0.026) | (0.028) |
| Pivotal Fair | | | | | 0.150 | | | | | | | | | 0.015 | | | | 0.045 | 0.036 | 0.052 |
| | | | | | (0.082) | | | | | | | | | (0.068) | | | | (0.050) | (0.052) | (0.046) |
| BF Resp (F) | | | | | | 0.379 | | | | | | | | | -0.192 | -0.280 | -0.275 | -0.190 | -0.276 | -0.226 |
| | | | | | | (0.188) | | | | | | | | | (0.201) | (0.209) | (0.231) | (0.134) | (0.153) | (0.167) |
| EA Engl (U) | | | | | | | -0.430*** | | | | | | | | | -0.283*** | | 0.093* | | |
| | | | | | | | (0.085) | | | | | | | | | (0.072) | | (0.037) | | |
| EA Engl (F) | | | | | | | | 0.416*** | | | | | | | 0.665*** | 0.539*** | | 0.177 | 0.191 | |
| | | | | | | | | (0.101) | | | | | | | (0.139) | (0.118) | | (0.108) | (0.110) | |
| EA Engl (F-U) | | | | | | | | | 0.283*** | | | | | | | | 0.408*** | | | -0.002 |
| | | | | | | | | | (0.060) | | | | | | | | (0.090) | | | (0.027) |
| EP Engl (U) | | | | | | | | | | -0.223*** | | | | | | -0.022 | | -0.175** | | |
| | | | | | | | | | | (0.049) | | | | | | (0.018) | | (0.056) | | |
| EP Engl (F) | | | | | | | | | | | 0.172 | | | | -0.156* | -0.137* | | 0.011 | 0.017 | |
| | | | | | | | | | | | (0.092) | | | | (0.065) | (0.064) | | (0.072) | (0.070) | |
| EP Engl (F-U) | | | | | | | | | | | | 0.137** | | | | | -0.054* | | | 0.116** |
| | | | | | | | | | | | | (0.044) | | | | | (0.022) | | | (0.034) |
| Engl Comb (F) | | | | | | | | | | | | | 0.306** | | | | | | | |
| | | | | | | | | | | | | | (0.106) | | | | | | | |
| Constant | 0.020 | 0.166*** | 0.072** | 0.165*** | 0.159*** | 0.136*** | 0.270*** | 0.080** | 0.173*** | 0.228*** | 0.138*** | 0.179*** | 0.107*** | 0.055 | 0.075** | 0.177*** | 0.199*** | 0.101** | 0.040 | 0.118** |
| | (0.031) | (0.025) | (0.023) | (0.005) | (0.008) | (0.019) | (0.019) | (0.023) | (0.000) | (0.012) | (0.019) | (0.001) | (0.023) | (0.030) | (0.021) | (0.025) | (0.023) | (0.032) | (0.027) | (0.037) |
| Observations | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 |
| R-squared | 0.062 | 0.000 | 0.038 | 0.002 | 0.005 | 0.010 | 0.033 | 0.030 | 0.042 | 0.015 | 0.008 | 0.016 | 0.018 | 0.070 | 0.036 | 0.048 | 0.047 | 0.073 | 0.071 | 0.072 |
| Number of Subjects | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 |

*Note:* Fixed effects regression with reward points as dependent variable (ranges from 0 to 7). *BF Responsibility Fair* represents a voters' share in the probability increase of a fair outcome. *Ex-ante (EA) and Ex-post (EP) Engl Responsibility (Un)fair* represent a voter's causal responsibility for a (un)fair event. We report within R-squared in the table. Robust standard errors in parentheses clustered at the subject level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

### 1.6.6 Econometric Comparison of Sanctioning Motives Variants

Table A1.24: Joint OLS regressions to compare the impact of the criteria on the usage of punishment and reward points

| | Punishment Points | | | Reward Points | |
|---|---|---|---|---|---|
| | Punishment | Both | | Reward | Both |
| Choice Unfair | 0.760*** | 0.486*** | Choice Fair | 0.460*** | 0.354*** |
| | (0.161) | (0.099) | | (0.115) | (0.082) |
| Outcome Unfair | -0.119 | -0.116* | Outcome Fair | -0.015 | -0.159** |
| | (0.109) | (0.053) | | (0.077) | (0.049) |
| Intention Unkind | 0.074 | 0.034 | Intention Kind | 0.176* | -0.066 |
| | (0.152) | (0.111) | | (0.082) | (0.053) |
| Initiator Unfair | 0.145 | 0.288 | Initiator Fair | -0.047 | -0.015 |
| | (0.230) | (0.242) | | (0.057) | (0.025) |
| Pivotal Unfair | 0.598** | 0.161 | Pivotal Fair | 0.267** | 0.045 |
| | (0.208) | (0.082) | | (0.080) | (0.050) |
| BF Responsibility (U) | 0.159 | -0.232 | BF Responsibility (F) | -0.042 | -0.244 |
| | (0.331) | (0.235) | | (0.268) | (0.128) |
| Ex-ante Engl Resp (U) | -0.411 | 0.092 | Ex-ante Engl Resp (U) | -0.089 | 0.208* |
| | (0.202) | (0.178) | | (0.146) | (0.100) |
| Ex-ante Engl Resp (F) | -0.087 | -0.037 | Ex-ante Engl Resp (F) | -0.079 | 0.107** |
| | (0.049) | (0.053) | | (0.072) | (0.039) |
| Ex-post Engl Resp (U) | 0.465** | 0.339*** | Ex-post Engl Resp (U) | 0.331** | 0.023 |
| | (0.152) | (0.075) | | (0.096) | (0.071) |
| Ex-post Engl Resp (F) | -0.033 | -0.095 | Ex-post Engl Resp (F) | 0.009 | -0.171** |
| | (0.114) | (0.067) | | (0.127) | (0.057) |
| Size of Majority | -0.022 | -0.043* | Size of Majority | -0.029 | 0.021 |
| | (0.026) | (0.019) | | (0.030) | (0.013) |
| Constant | 0.196 | 0.245** | Constant | 0.267* | 0.0184 |
| | (0.130) | (0.076) | | (0.123) | (0.066) |
| Observations | 9,600 | 9,600 | Observations | 9,280 | 9,600 |
| R-squared | 0.211 | 0.152 | R-squared | 0.164 | 0.074 |
| Number of Subjects | 30 | 30 | Number of Subjects | 29 | 30 |

Note: OLS fixed effects regressions with punishment points and reward points as dependent variables. Punishment points (left side of the table) can take values from 0 to 7 and are used in the treatments *Punishment* and *Both*. Reward points (right side of the table) can take values from 0 to 7 and are used in the treatments *Reward* and *Both*. *Choice (Un)fair* equals 1 if the (un)fair allocation is chosen. *Outcome (Un)fair* is a dummy that equals 1 if the (un)fair outcome is implemented. *Intention (Un)kind* equals 1 if a voter votes for the (un)fair allocation while no majority was reached before. *Initiator (Un)fair* equals 1 if a voter is the initiator for the (un)fair outcome. *Pivotal (Un)fair* is an indicator that equals 1 if a voter is pivotal for the (un)fair outcome. *BF Responsibility (Un)fair* and *Ex-ante and Ex-post Engl Responsibility (Un)fair* correspond to the responsibility measures explained in Section 1.3. *Size of Majority* indicates the number of majority voters and can take values from 3 to 5. Robust standard errors in parentheses clustered at the subject level.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A1.25: Joint OLS regressions to compare the impact of the criteria on the usage of punishment and reward points

| | Punishment Points | | | Reward Points | |
|---|---|---|---|---|---|
| | Punishment | Both | | Reward | Both |
| Choice Unfair | 0.803*** | 0.527*** | Choice Fair | 0.487*** | 0.371*** |
| | (0.178) | (0.104) | | (0.125) | (0.087) |
| Outcome Unfair | -0.090 | -0.071* | Outcome Fair | -0.011 | -0.088** |
| | (0.061) | (0.033) | | (0.042) | (0.026) |
| Intention Unkind | 0.073 | 0.038 | Intention Kind | 0.177* | -0.073 |
| | (0.152) | (0.111) | | (0.082) | (0.055) |
| Initiator Unfair | 0.144 | 0.280 | Initiator Fair | -0.051 | -0.024 |
| | (0.231) | (0.242) | | (0.062) | (0.025) |
| Pivotal Unfair | 0.600** | 0.175* | Pivotal Fair | 0.270** | 0.037 |
| | (0.208) | (0.085) | | (0.080) | (0.052) |
| BF Responsibility (U) | 0.111 | -0.340 | BF Responsibility (F) | -0.051 | -0.339* |
| | (0.399) | (0.265) | | (0.326) | (0.151) |
| Ex-ante Engl Resp (U) | -0.389 | 0.103 | Ex-ante Engl Resp (F) | -0.084 | 0.228* |
| | (0.212) | (0.180) | | (0.150) | (0.105) |
| Ex-post Engl Resp (U) | 0.454** | 0.346*** | Ex-post Engl Resp (F) | 0.327** | 0.034 |
| | (0.148) | (0.072) | | (0.093) | (0.068) |
| Size of Majority | -0.013 | -0.033* | Size of Majority | -0.023 | 0.026* |
| | (0.023) | (0.016) | | (0.026) | (0.011) |
| Constant | 0.106 | 0.143* | Constant | 0.215* | -0.052 |
| | (0.093) | (0.055) | | (0.094) | (0.051) |
| Observations | 9,600 | 9,600 | Observations | 9,280 | 9,600 |
| $R^2$ | 0.211 | 0.152 | $R^2$ | 0.164 | 0.072 |
| Number of Subjects | 30 | 30 | Number of Subjects | 29 | 30 |

Note: OLS fixed effects regressions with punishment points and reward points as dependent variables. Punishment points (left side of the table) can take values from 0 to 7 and are used in the treatments *Punishment* and *Both*. Reward points (right side of the table) can take values from 0 to 7 and are used in the treatments *Reward* and *Both*. *Choice (Un)fair* equals 1 if the (un)fair allocation is chosen. *Outcome (Un)fair* is a dummy that equals 1 if the (un)fair outcome is implemented. *Intention (Un)kind* equals 1 if a voter votes for the (un)fair allocation while no majority was reached before. *Initiator (Un)fair* equals 1 if a voter is the initiator for the (un)fair outcome. *Pivotal (Un)fair* is an indicator that equals 1 if a voter is pivotal for the (un)fair outcome. *BF Responsibility (Un)fair* and *Ex-ante and Ex-post Engl Responsibility (Un)fair* correspond to the responsibility measures explained in Section 1.3. *Size of Majority* indicates the number of majority voters and can take values from 3 to 5. Robust standard errors in parentheses clustered at the subject level.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

### 1.6.7 Econometric Comparison of Sanctioning Motives Conditional on Choice

Table A1.26: Econometric Comparison of Sanctioning Motives Conditional on Choice

| | Choice Unfair | | | | Choice Fair | | |
|---|---|---|---|---|---|---|---|
| | Punishment | Both | Reward | | Reward | Both | Punishment |
| | Punishment Points | | Reward Points | | Reward Points | | Punishment Points |
| Outcome Unfair | 0.112 | 0.0104 | -0.076 | Outcome Fair | 0.145 | -0.325* | -0.001 |
| | (0.239) | (0.122) | (0.065) | | (0.170) | (0.120) | (0.015) |
| Intention Unkind | 0.0216 | 0.126 | -0.033 | Intention Kind | 0.216* | 0.016 | 0.008 |
| | (0.132) | (0.069) | (0.020) | | (0.084) | (0.043) | (0.012) |
| Initiator Unfair | 0.264 | 0.357 | 0.016 | Initiator Fair | 0.004 | 0.060 | -0.011 |
| | (0.227) | (0.241) | (0.019) | | (0.051) | (0.035) | (0.014) |
| Pivotal Unfair | 0.611** | 0.162 | 0.022 | Pivotal Fair | 0.260** | 0.115* | -0.003 |
| | (0.205) | (0.112) | (0.017) | | (0.084) | (0.043) | (0.010) |
| Size of Majority | -0.064 | -0.106* | 0.011 | Size of Majority | -0.095 | 0.089* | 0.011 |
| | (0.052) | (0.041) | (0.031) | | (0.064) | (0.040) | (0.019) |
| Constant | 1.002*** | 0.904*** | 0.178* | Constant | 0.898*** | 0.196 | -0.003 |
| | (0.254) | (0.137) | (0.081) | | (0.234) | (0.127) | (0.066) |
| Observations | 4,800 | 4,800 | 4,640 | Observations | 4,640 | 4,800 | 4,800 |
| R-squared | 0.038 | 0.023 | 0.004 | R-squared | 0.031 | 0.026 | 0.001 |
| Number of Subjects | 30 | 30 | 29 | Number of Subjects | 29 | 30 | 30 |

*Note:* OLS fixed effects regression with punishment points and reward points as dependent variables (ranges from 0 to 7) conditional on unfair and fair choice. Robust standard errors in parentheses clustered at the subject level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

### 1.6.8 Hurdle Model

In this section, we present the results of a hurdle model. The model consists of two stages through which we estimate the intensive and extensive margin of each predictor. In the first stage, we estimate which predictors explain the decision to sanction and in the second stage we analyze which variables impact the amount of sanctioning. Note that we use a slightly modified hurdle model. In the second stage, we consider all the sanctioning decisions where at least one voter received sanction points and not only the decision in which the sanction is positive. Therefore, the second stage measures how sanctions are allocated between the different voters in the situation in which at least one voter is sanctioned. We use this procedure because in our experiment the number of total sanctioning points was restricted.

Table A1.27: Punishment Treatment - Marginal Effects Probit Regression (Hurdle Model)

| Prob. Sanction | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 | Model 7 |
|---|---|---|---|---|---|---|---|
| Choice Unfair | 0.400*** | | | | | | 0.327*** |
| | (0.046) | | | | | | (0.049) |
| Outcome Unfair | | 0.262*** | | | | | 0.098** |
| | | (0.044) | | | | | (0.036) |
| Intention Unkind | | | 0.365*** | | | | 0.022 |
| | | | (0.044) | | | | (0.020) |
| Initiator Unfair | | | | 0.311*** | | | 0.022 |
| | | | | (0.054) | | | (0.018) |
| Pivotal Unfair | | | | | 0.341*** | | 0.038* |
| | | | | | (0.054) | | (0.018) |
| Size of Majority | | | | | | 0.038** | 0.029 |
| | | | | | | (0.012) | (0.015) |
| Observations | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 |
| Pseudo R-squared | 0.261 | 0.100 | 0.183 | 0.041 | 0.049 | 0.003 | 0.290 |

*Note:* Marginal effects from probit regression with probability to sanction as dependent variable. Robust standard errors in parentheses clustered at the subject level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A1.28: Reward Treatment - Marginal Effects Probit Regression (Hurdle Model)

| Prob. Sanction | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 | Model 7 |
|---|---|---|---|---|---|---|---|
| Choice Fair | 0.402*** | | | | | | 0.294*** |
| | (0.063) | | | | | | (0.063) |
| Outcome Fair | | 0.268*** | | | | | 0.116** |
| | | (0.044) | | | | | (0.036) |
| Intention Kind | | | 0.384*** | | | | 0.069* |
| | | | (0.060) | | | | (0.029) |
| Initiator Fair | | | | 0.33*** | | | 0.045* |
| | | | | (0.049) | | | (0.018) |
| Pivotal Fair | | | | | 0.331*** | | 0.043* |
| | | | | | (0.052) | | (0.019) |
| Size of Majority | | | | | | 0.016 | 0.017 |
| | | | | | | (0.018) | (0.019) |
| Observations | 9,280 | 9,280 | 9,280 | 9,280 | 9,280 | 9,280 | 9,280 |
| Pseudo R-squared | 0.192 | 0.081 | 0.157 | 0.039 | 0.038 | 0.001 | 0.215 |

*Note:* Marginal effects from probit regression with probability to sanction as dependent variable. Robust standard errors in parentheses clustered at the subject level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A1.29: Both Treatment Unfair Variables - Marginal Effects Probit Regression (Hurdle Model)

| Prob. Sanction | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 | Model 7 |
|---|---|---|---|---|---|---|---|
| Choice Unfair | 0.322*** | | | | | | 0.266*** |
| | (0.049) | | | | | | (0.045) |
| Outcome Unfair | | 0.187*** | | | | | 0.054** |
| | | (0.029) | | | | | (0.017) |
| Intention Unkind | | | 0.301*** | | | | 0.022 |
| | | | (0.046) | | | | (0.013) |
| Initiator Unfair | | | | 0.269*** | | | 0.022 |
| | | | | (0.047) | | | (0.018) |
| Pivotal Unfair | | | | | 0.231*** | | 0.007 |
| | | | | | (0.039) | | (0.009) |
| Size of Majority | | | | | | 0.002 | -0.002 |
| | | | | | | (0.009) | (0.009) |
| Observations | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 |
| Pseudo R-squared | 0.236 | 0.070 | 0.171 | 0.040 | 0.030 | 0.00 | 0.250 |

*Note:* Marginal effects from probit regression with probability to sanction as dependent variable. Robust standard errors in parentheses clustered at the subject level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A1.30: Both Treatment Fair Variables - Marginal Effects Probit Regression (Hurdle Model)

| Prob. Sanction | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 | Model 7 |
|---|---|---|---|---|---|---|---|
| Choice Fair | 0.175*** | | | | | | 0.170*** |
| | (0.031) | | | | | | (0.031) |
| Outcome Fair | | 0.046 | | | | | -0.022 |
| | | (0.024) | | | | | (0.015) |
| Intention Kind | | | 0.153*** | | | | 0.014* |
| | | | (0.028) | | | | (0.007) |
| Initiator Fair | | | | 0.091** | | | 0.003 |
| | | | | (0.030) | | | (0.006) |
| Pivotal Fair | | | | | 0.083** | | -0.001 |
| | | | | | (0.027) | | (0.005) |
| Size of Majority | | | | | | 0.008 | 0.011* |
| | | | | | | (0.008) | (0.006) |
| Observations | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 | 9,600 |
| Pseudo R-squared | 0.160 | 0.010 | 0.100 | 0.011 | 0.010 | 0.001 | 0.163 |

*Note:* Marginal effects from probit regression with probability to sanction as dependent variable. Robust standard errors in parentheses clustered at the subject level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A1.31: Punishment Treatment - Linear Regression Conditional on Punishment Points (Hurdle Model)

| Punishment Points | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 | Model 7 |
|---|---|---|---|---|---|---|---|
| Choice Unfair | 1.872*** | | | | | | 1.805*** |
| | (0.135) | | | | | | (0.273) |
| Outcome Unfair | | 0.225** | | | | | -0.620*** |
| | | (0.0685) | | | | | (0.114) |
| Intention Unkind | | | 1.649*** | | | | 0.148 |
| | | | (0.124) | | | | (0.215) |
| Initiator Unfair | | | | 0.747* | | | 0.116 |
| | | | | (0.342) | | | (0.340) |
| Pivotal Unfair | | | | | 1.377*** | | 0.656* |
| | | | | | (0.299) | | (0.280) |
| Size of Majority | | | | | | -0.0269 | -0.245*** |
| | | | | | | (0.017) | (0.036) |
| Constant | 0.0996** | 1.036*** | 0.387*** | 1.089*** | 1*** | 1.286*** | 1.235*** |
| | (0.031) | (0.091) | (0.044) | (0.071) | (0.059) | (0.075) | (0.130) |
| Observations | 4,375 | 4,375 | 4,375 | 4,375 | 4,375 | 4,375 | 4,375 |
| R-squared | 0.333 | 0.004 | 0.266 | 0.026 | 0.090 | 0.000 | 0.382 |

*Note:* Linear regression with punishment points as dependent variable (ranges from 0 to 7). Robust standard errors in parentheses clustered at the subject level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A1.32: Reward Treatment - Linear Regression Conditional on Reward Points (Hurdle Model)

| Reward Points | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 | Model 7 |
|---|---|---|---|---|---|---|---|
| Choice Fair | 1.248*** | | | | | | 0.999*** |
| | (0.228) | | | | | | (0.251) |
| Outcome Fair | | 0.229** | | | | | -0.292** |
| | | (0.076) | | | | | (0.085) |
| Intention Kind | | | 1.171*** | | | | 0.395** |
| | | | (0.207) | | | | (0.136) |
| Initiator Fair | | | | 0.454*** | | | -0.134 |
| | | | | (0.100) | | | (0.102) |
| Pivotal Fair | | | | | 0.896*** | | 0.251 |
| | | | | | (0.185) | | (0.167) |
| Size of Majority | | | | | | -0.0316 | -0.105** |
| | | | | | | (0.026) | (0.036) |
| Constant | 0.313* | 0.861*** | 0.457*** | 0.950*** | 0.893*** | 1.116*** | 0.797*** |
| | (0.130) | (0.092) | (0.102) | (0.069) | (0.064) | (0.140) | (0.137) |
| Observations | 4,805 | 4,805 | 4,805 | 4,805 | 4,805 | 4,805 | 4,805 |
| R-squared | 0.233 | 0.007 | 0.207 | 0.014 | 0.055 | 0.000 | 0.259 |

*Note:* Linear regression with reward points as dependent variable (ranges from 0 to 7). Robust standard errors in parentheses clustered at the subject level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A1.33: Both Treatment - Linear Regression Conditional on Punishment Points (Hurdle Model)

| Punishment Points | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 | Model 7 |
|---|---|---|---|---|---|---|---|
| Choice Unfair | 1.730*** | | | | | | 1.512*** |
| | (0.104) | | | | | | (0.172) |
| Outcome Unfair | | 0.252*** | | | | | -0.476*** |
| | | (0.055) | | | | | (0.074) |
| Intention Unkind | | | 1.576*** | | | | 0.347* |
| | | | (0.114) | | | | (0.144) |
| Initiator Unfair | | | | 1.178** | | | 0.388 |
| | | | | (0.413) | | | (0.450) |
| Pivotal Unfair | | | | | 0.726* | | -0.00313 |
| | | | | | (0.263) | | (0.231) |
| Size of Majority | | | | | | -0.009 | -0.150*** |
| | | | | | | (0.030) | (0.038) |
| Constant | 0.0833** | 0.889*** | 0.302*** | 0.900*** | 0.961*** | 1.090*** | 0.816*** |
| | (0.029) | (0.058) | (0.031) | (0.061) | (0.061) | (0.124) | (0.143) |
| Observations | 3,550 | 3,550 | 3,550 | 3,550 | 3,550 | 3,550 | 3,550 |
| R-squared | 0.324 | 0.006 | 0.273 | 0.071 | 0.027 | 0.000 | 0.356 |

*Note:* Linear regression with punishment points as dependent variable (ranges from 0 to 7). Robust standard errors in parentheses clustered at the subject level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A1.34: Both Treatment - Linear Regression Conditional on Reward Points (Hurdle Model)

| Reward Points | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 | Model 7 |
|---|---|---|---|---|---|---|---|
| Choice Fair | 1.313*** | | | | | | 1.489*** |
| | (0.117) | | | | | | (0.212) |
| Outcome Fair | | 0.157 | | | | | -0.390*** |
| | | (0.084) | | | | | (0.083) |
| Intention Kind | | | 1.075*** | | | | -0.0335 |
| | | | (0.105) | | | | (0.151) |
| Initiator Fair | | | | 0.451** | | | -0.126 |
| | | | | (0.127) | | | (0.102) |
| Pivotal Fair | | | | | 0.715** | | 0.113 |
| | | | | | (0.247) | | (0.230) |
| Size of Majority | | | | | | 0.050 | 0.024 |
| | | | | | | (0.041) | (0.042) |
| Constant | 0.079 | 0.647*** | 0.292*** | 0.678*** | 0.653*** | 0.548** | 0.108 |
| | (0.039) | (0.059) | (0.041) | (0.055) | (0.056) | (0.159) | (0.164) |
| Observations | 2,320 | 2,320 | 2,320 | 2,320 | 2,320 | 2,320 | 2,320 |
| R-squared | 0.322 | 0.005 | 0.207 | 0.013 | 0.032 | 0.001 | 0.349 |

*Note:* Linear regression with reward points as dependent variable (ranges from 0 to 7). Robust standard errors in parentheses clustered at the subject level. $^{*}$ $p < 0.05$, $^{**}$ $p < 0.01$, $^{***}$ $p < 0.001$

### 1.6.9 Time Trends: Econometric Comparison of Sanctioning Motives

Table A1.35: Time Trends - Econometric Comparison of Punishing Motives

| | Decision No. 1 - 21 | | Decision No. 22 - 42 | | Decision No. 43 - 64 | |
|---|---|---|---|---|---|---|
| Punishment Points | Punishment | Both | Punishment | Both | Punishment | Both |
| Choice Unfair | 0.641*** | 0.570*** | 0.868*** | 0.500** | 0.823*** | 0.411** |
| | (0.148) | (0.120) | (0.223) | (0.140) | (0.224) | (0.136) |
| Outcome Unfair | -0.013 | -0.004 | 0.068 | -0.010 | 0.054 | 0.013 |
| | (0.102) | (0.063) | (0.107) | (0.0721) | (0.109) | (0.075) |
| Intention Unkind | 0.281* | 0.087 | -0.078 | 0.244* | -0.075 | 0.144 |
| | (0.133) | (0.101) | (0.179) | (0.095) | (0.192) | (0.090) |
| Initiator Unfair | 0.220 | 0.403 | 0.233 | 0.342 | 0.434 | 0.316 |
| | (0.194) | (0.205) | (0.225) | (0.285) | (0.316) | (0.256) |
| Pivotal Unfair | 0.560* | 0.300 | 0.628** | -0.059 | 0.742** | 0.226 |
| | (0.228) | (0.163) | (0.227) | (0.163) | (0.247) | (0.113) |
| Size of Majority | 0.021 | -0.092** | -0.023 | -0.007 | -0.032 | -0.058 |
| | (0.031) | (0.029) | (0.040) | (0.045) | (0.047) | (0.035) |
| Constant | -0.019 | 0.351** | 0.095 | 0.0444 | 0.117 | 0.231 |
| | (0.121) | (0.104) | (0.156) | (0.149) | (0.171) | (0.128) |
| Observations | 3,150 | 3,150 | 3150 | 3150 | 3,300 | 3,300 |
| R-squared | 0.207 | 0.177 | 0.203 | 0.147 | 0.217 | 0.132 |
| Number of Subjects | 30 | 30 | 30 | 30 | 30 | 30 |

*Note:* OLS fixed effects regression with punishment points as dependent variable (ranges from 0 to 7). Robust standard errors in parentheses clustered at the subject level.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A1.36: Time Trends - Econometric Comparison of Rewarding Motives

| | Decision No. 1 - 21 | | Decision No. 22 - 42 | | Decision No. 43 - 64 | |
|---|---|---|---|---|---|---|
| Reward Points | Reward | Both | Reward | Both | Reward | Both |
| Choice Fair | 0.597*** | 0.373*** | 0.340* | 0.353** | 0.453** | 0.376** |
| | (0.155) | (0.088) | (0.133) | (0.121) | (0.155) | (0.126) |
| Outcome Fair | 0.098 | -0.111** | 0.115 | -0.112* | 0.041 | -0.103 |
| | (0.073) | (0.038) | (0.080) | (0.046) | (0.075) | (0.058) |
| Intention Kind | 0.118 | -0.006 | 0.405** | -0.041 | 0.209 | 0.0118 |
| | (0.120) | (0.056) | (0.116) | (0.101) | (0.110) | (0.068) |
| Initiator Fair | 0.043 | -0.0215 | 0.007 | 0.002 | 0.042 | -0.110* |
| | (0.079) | (0.052) | (0.084) | (0.054) | (0.065) | (0.042) |
| Pivotal Fair | 0.235 | -0.046 | 0.213 | 0.053 | 0.395* | 0.030 |
| | (0.120) | (0.059) | (0.125) | (0.086) | (0.173) | (0.091) |
| Size of Majority | 0.008 | 0.018 | -0.040 | 0.022 | -0.046 | -3.00e-05 |
| | (0.040) | (0.027) | (0.038) | (0.022) | (0.042) | (0.022) |
| Constant | 0.109 | 0.007 | 0.253 | -0.014 | 0.263 | 0.027 |
| | (0.125) | (0.110) | (0.146) | (0.098) | (0.145) | (0.076) |
| Observations | 3,045 | 3,150 | 3,045 | 3,150 | 3,190 | 3,300 |
| R-squared | 0.151 | 0.079 | 0.173 | 0.061 | 0.171 | 0.078 |
| Number of Subjects | 29 | 30 | 29 | 30 | 29 | 30 |

*Note:* OLS fixed effects regression with reward points as dependent variable (ranges from 0 to 7). Robust standard errors in parentheses clustered at the subject level.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

### 1.6.10 Finite Mixture Model Analysis

Table A1.37: Parameter Sets for each Component of Finite Mixture Models

| | Punishment | | | Both (Punishment Points) | | |
|---|---|---|---|---|---|---|
| | Cluster 1 Little | Cluster 2 Pivotal | Cluster 3 Choice | Cluster 1 No | Cluster 2 Little | Cluster 3 Choice / Intention |
| Choice Unfair | 0.111 | 0.505 | 1.961 | 0.020 | 0.457 | 0.767 |
| Outcome Unfair | 0.092 | 0.409 | -0.642 | 0.010 | 0.032 | -0.010 |
| Intention Unkind | 0.122 | 0.285 | -0.436 | 0.014 | -0.121 | 0.396 |
| Initiator Unfair | -0.080 | 0.244 | 0.812 | 0.022 | 0.001 | 0.741 |
| Pivotal Unfair | -0.021 | 1.355 | 0.242 | -0.000 | 0.088 | 0.278 |
| Size of Majority | 0.063 | 0.000 | -0.126 | -0.004 | 0.024 | -0.125 |
| Constant | -0.174 | -0.115 | 0.670 | 0.011 | -0.054 | 0.468 |
| $\sigma$ | 0.634 | 0.895 | 1.406 | 0.216 | 0.635 | 1.178 |
| Number of Subjects | 9 | 13 | 8 | 7 | 9 | 14 |
| | Reward | | | Both (Reward Points) | | |
| | Cluster 1 Little | Cluster 2 Pivotal | Cluster 3 Choice | Cluster 1 No | Cluster 2 Little | Cluster 3 Choice /Intention |
| Choice Fair | 0.351 | 0.435 | 0.758 | 0.000 | 0.290 | 0.844 |
| Outcome Fair | 0.087 | 0.194 | -0.254 | -0.000 | -0.088 | -0.253 |
| Intention Kind | -0.025 | 0.433 | 0.266 | 0.000 | -0.003 | -0.055 |
| Initiator Fair | 0.121 | 0.086 | -0.312 | 0.000 | -0.052 | -0.056 |
| Pivotal Fair | 0.080 | 0.584 | -0.137 | 0 .000 | 0.002 | 0.045 |
| Size of Majority | 0.051 | -0.036 | -0.177 | 0.000 | 0.018 | 0.010 |
| Constant | -0.179 | 0.131 | 1.281 | -0.000 | -0.031 | 0.113 |
| $\sigma$ | 0.476 | 0.899 | 1.458 | 0.006 | 0.481 | 0.936 |
| Number of Subjects | 10 | 14 | 5 | 7 | 15 | 8 |

Note: The table reports the parameter sets for each component resulting from the finite mixture models across treatments, including the estimated variance $\sigma$ and the number of subjects classified into the respective component. The finite mixture models are estimated via a general linear regression using an EM-algorithm and the number of components was determined by the best goodness of fit.

### 1.6.11 Eye–Tracking Results

We collected the data for 90 participants. Three subjects had to be excluded due to poor gaze data quality. The recipients made choices on two decision screens. On the first screen, the recipients made the decision whether to sanction or not. In addition to the votes being displayed, the decision screen had two large bars at the bottom indicating the willingness to sanction or not (see Figure A1.2 in Appendix 1.6.1). On the second screen, the recipients decided on how many sanction points they wanted to allocate to the voters.[25] The second decision screen showed the votes, as well as additional buttons for allocating punishment and reward points (see Figure 1.1). Since on the second decision screen, subjects could allot the sanction points to each voter, it involved a lot of clicking and focus on the buttons. Therefore, we briefly discuss the results for the first decision screen.

To examine whether the gaze data of the recipients is in line with their sanctioning behavior, we use the average number of fixations and the dwell time. Generally, the fixations on different pieces of information are related to the processing of the inspected information (Just and Carpenter, 1980) and indicate the relative importance of specific information for the decision-making process (Duchowski, 2017; Rahal and Fiedler, 2019).

It is important to note that certain voter positions can attract relatively more fixations merely due to being at the center of the screen. Also, scenarios in which a voter is salient might attract more attention. For instance, in situations where one voter votes differently than the other four voters, saliency is strong and might influence the gaze pattern of the recipients. In order to understand the importance of positioning and saliency, we conduct a fixed effects regression with the share of fixations as a dependent variable fixing on Voter Position 1. We control for voter position and saliency in this regression, and the results are shown in Table A1.38.

Compared to the share of fixations on Voter 1, voters on position 2 and 3 receive a higher share of fixations, while voters on position 5 receive a lower share of fixations. This can be due to the natural reading habit of reading from left to right in Western culture. Also, voters who are part of the minority group receive more fixations due to their saliency. This effect is even stronger when there is only one minority voter. The regression output shows that both saliency and voter position highly influence the gaze data.

[25] Only the recipients who decided to sanction moved on to the second screen.

Table A1.38: Share of Fixations - Importance of Position and Saliency

| | Punishment | Reward | Both |
|---|---|---|---|
| Voter Position 2 | 0.036** | 0.047** | 0.024 |
| | (0.012) | (0.014) | (0.017) |
| Voter Position 3 | 0.130*** | 0.141*** | 0.123*** |
| | (0.023) | (0.031) | (0.029) |
| Voter Position 4 | -0.003 | -0.002 | -0.023 |
| | (0.014) | (0.014) | (0.017) |
| Voter Position 5 | -0.039* | -0.054*** | -0.083*** |
| | (0.018) | (0.014) | (0.017) |
| Only One Salient | 0.053*** | 0.068*** | 0.071*** |
| | (0.009) | (0.012) | (0.009) |
| Neighbor of Only One Salient | 0.005 | 0.009 | -0.003 |
| | (0.003) | (0.005) | (0.004) |
| Minority | 0.010* | 0.017* | 0.008 |
| | (0.004) | (0.006) | (0.004) |
| Constant | 0.168*** | 0.163*** | 0.185*** |
| | (0.012) | (0.013) | (0.015) |
| Observations | 9,295 | 8,695 | 8,925 |
| R-squared | 0.151 | 0.169 | 0.179 |
| Number of Subjects | 30 | 28 | 29 |

Note: Fixed effects regression with share of fixations as dependent variable fixing on Voter Position 1. *Voter Position 2-5* indicate the position of the voter in the decision scenario. *Only One Salient* is a dummy variable indicating if the voter is the only salient voter. *Neighbor of Only One Salient* is a dummy variable that takes the value 1 if the voter is next to the only salient voter. *Minority* is a dummy variable indicating if the voter is part of the minority.
Robust standard errors in parentheses clustered at the subject level.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

We will now discuss how the individual sanction motives (discussed in Section 1.3) impact the fixations. We regress the share of fixations of the recipients on the different individual sanction motives incorporated in each decision screen. In addition, we control for the position and saliency measures mentioned above. The regression output is shown in Table A1.39 and suggests that voters are focused more when being the initiator. The pivotal voter does not receive a special focus. The result is against our expectations, as the pivotal voter is punished and rewarded the most. A potential reason could be that it is much more difficult to assess the responsibility of the initiator compared to the pivotal voter and, thus, the recipients look longer at the initiator.

Table A1.39: Share of Fixations - Impact of Sanction Motives

| | Punishment | Both | | Reward | Both |
|---|---|---|---|---|---|
| Choice Unfair | -0.001 | 0.005 | Choice Fair | 0.010 | -0.005 |
| | (0.006) | (0.005) | | (0.006) | (0.004) |
| Intention Unkind | 0.004 | -0.0002 | Intention Kind | -0.007 | -0.007 |
| | (0.007) | (0.006) | | (0.007) | (0.006) |
| Initiator Unfair | 0.017* | 0.021* | Initiator Fair | 0.024** | 0.016* |
| | (0.007) | (0.008) | | (0.009) | (0.006) |
| Pivotal Unfair | -0.005 | -0.005 | Pivotal Fair | 0.006 | 0.001 |
| | (0.007) | (0.006) | | (0.011) | (0.009) |
| Controls | Yes | Yes | Controls | Yes | Yes |
| Constant | 0.160*** | 0.175*** | Constant | 0.152*** | 0.185*** |
| | (0.011) | (0.015) | | (0.012) | (0.015) |
| Observations | 9,295 | 8,925 | Observations | 8,695 | 8,925 |
| R-squared | 0.152 | 0.181 | R-squared | 0.171 | 0.180 |
| Number of Subjects | 30 | 29 | Number of Subjects | 28 | 29 |

Note: OLS fixed effects regression with share of fixation of the first screen as the dependent variable fixing on Voter Position 1. The controls include variables for *voter position 2-5, saliency and minority.*
Robust standard errors in parentheses clustered at the subject level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

As the recipients allocate sanction points to individual voters on the second decision screen, it is also possible that they might fixate more on the different voters while sanctioning. However, on the second screen, the measures for share of fixations and dwell time show a similar trend as on the first screen. Voter position and saliency strongly influence the gaze pattern of the recipients, even on the second decision screen. With respect to the sanction motives, choosing an unfair allocation as well as initiating an unfair outcome leads to receiving a higher share of fixations as compared to the other sanction motives in the *Punishment* and *Both* treatments. In the *Reward* treatment, initiating a fair outcome and intentionally voting for a fair outcome attracts a higher share of fixations. Overall, the gaze pattern on the first and second screen exhibited similar trends.

# Chapter 2

# Using Mental Imagery to Foster Future-Mindedness in Long-Term Decisions

Regina Stumpf[1,2] and Baiba Renerte[3]

[1] Department of Economics, University of Konstanz (Germany)
[2] Thurgau Institute of Economics (Switzerland)
[3] Department of Finance, University of Zurich (Switzerland)

**Abstract**

What encourages individuals and groups to envision the long-term future and act on that basis? We study how decisions can be shifted from short-term to long-term investment options. We conduct a laboratory experiment to test an intervention of mental imagery — creating visual images in the mind's eye — and compare it to other standard modes of decision processing. Subjects face multiple decisions to invest in a real-world short-term or long-term project. Our results show that mental imagery leads to more long-term decisions compared to other established decision-making mindsets. We propose a mechanism by which mental imagery fosters future-mindedness by influencing the perception of the time horizon, and show that this mechanism works particularly well for decision makers with optimistic inclinations about their own future life satisfaction. Collective long-term decisions are mainly determined by individual long-term preferences. Our project implies that future-mindedness can be fostered by using simple mental-imagery interventions at the individual-level.

**Keywords:** mental imagery, future-mindedness, long-term decision-making

**JEL Classification:** C91, C92, D64, D91

## 2.1 Introduction

*"If one is mentally out of breath all the time from dealing with the present, there is no energy left for imagining the future."*
*– Elise Boulding, sociologist (1978)*

The quote by Elise Boulding captures the inherent dynamics of balancing and prioritizing present and future problems. According to the Global Risks Report 2023 of the World Economic Forum, the climate crisis is one of the biggest long-term challenges. Yet, it is one the world is ill-prepared to tackle because of the short-term crises the society faces. To put it in the words of Elise Boulding: These short-term crises claim most of our current breath, such that there is no energy left for envisioning the future to solve long-term problems. Different disciplines capture the same obstacle under different terms, such as short-termism, presentism, present-bias and myopia (e.g., Olesiński et al. (2014); Busemeyer (2024)). They all define a bias in decision-making in favor of the short-term at the cost of the long-term. Short-termism is present in the private sector where it has increased over time within and across companies (Sampson and Shi, 2023), although firms with a long-term focus outperform those with a short-term vision with respect to growth and earnings (Barton et al., 2017). Companies often have short performance evaluation intervals – focusing on quarterly results – instead of generating a long-term value for the company (Olesiński et al., 2014; Davies et al., 2014). Several reasons are indicated, such as the pressure and expectations of shareholders, new technologies and the globalization of the financial market. Similarly, in politics the elected government often fails to invest in the present to tackle long-term problems such as a functioning pension system, a working infrastructure or the fight against climate change (Jacobs, 2016; Ferrera, 2017). Reasons are often the short-term incentives for politicians, short-term preferences of voters and interest groups, and the lack of representation of future generations (MacKenzie, 2016). Although political presentism is widely distributed among citizens, there is some heterogeneity in long-term policy preferences (Busemeyer, 2024). The laws of many societies show a clear preference for supporting elderly citizens at the cost of younger ones, since the former often represent a majority of the population (Thompson, 2010).

These problems have the common characteristic that the relevant decision makers will not be the ones to directly benefit from the realization of these projects, such that a long-term vision is crucial to successfully overcome the challenges of short-termism. Across different disciplines, the notion of a long-term vision, a preference for the future, a future-mindedness has been argued to help in overcoming short-termism (e.g., Ross et al. (2021); Reece et al. (2022)). The United Nations (UN) launched a variety of initiatives to enhance the ability of foresight and prospection. [1] Prospection is the ability of using the mental process of evaluating future possibilities and then using these projections for the guidance of thought and action in the present (Buckner and Carroll, 2007; Gilbert and Wilson, 2007). Arguably, humans already do very well in prospecting the future and have the longest time horizon of anticipation and

[1] `https://un-two-zero.network/foresight/`, last retrieved: 05.04.2024

mental time travel in comparison to other species (Seligman et al., 2013, 2016). Thinking about the future therefore influences our future-mindedness and starts by imagining future outcomes (Baumeister et al., 2016). In their book on prospection, Seligman et al. (2016) propose three categories of guidance for future thought: intuitive, deliberative and imaginative prospection. Similarly, other studies investigate the mechanism of spontaneous or deliberative future thinking (Cole and Kvavilashvili, 2021; Duffy and Cole, 2021).

In this paper, we are interested in answering the following question: What encourages individuals and groups to envision the long-term future and act on that basis? In other words, how to foster future-mindedness? We conduct a laboratory and online experiment to test a simple intervention of mental imagery – creating visual images in the mind's eye – and compare it to distinct decision framing interventions to examine whether it can foster future-mindedness and therefore encourage a shift from short-term to long-term outcomes. In a laboratory experiment, subjects face multiple real-world investment decisions, in which they choose to invest a given amount in either a short-term or a long-term project within the organization they are affiliated with. The short-term projects are realized within the next year, while the realization of the long-term projects takes several years. Along the lines of Seligman et al. (2016) we construct three treatment variations which put subjects into an imaginative, intuitive or deliberative mindset before making an investment decision. In addition, we apply these interventions not only in individual, but also collective decision-making settings. Our results show that mental imagery can successfully be implemented via short and simple instructions. Mental imagery leads to more long-term decisions compared to other established mindsets when deciding individually. When deciding in groups, the initial group composition determines the choice of the group, suggesting that individual preferences paired with mental imagery are important for shifting collective decisions towards the long-term.

We propose a mechanism of how mental imagery fosters future-mindedness by affecting the perception of the time horizon of long-term outcomes. Linked to Construal Level Theory, (Trope and Liberman, 2010; Liberman and Trope, 2003) mental imagery reduces the psychological distance to an object by thinking more concretely about it. This in turn leads to more long-term decisions. Our data can partially confirm this conceptual framework. Furthermore, we show that mental imagery leads to more long-term choices, especially for subjects with an optimistic view about their future.

We also test the implications of our experiment with a representative sample of the German working population in an online experiment. We show that our intervention of mental imagery can also be successfully implemented for a general sample in an online setting. The data of the online experiment weakly confirm that mental imagery leads to a shift towards future outcomes. This effect is driven by the vividness of the mental imagery intervention.

The economic literature on time preferences in intertemporal decisions has posed the question of how prosociality is influenced by delayed realizations in a charitable context (Breman, 2011; Andreoni and Serra-Garcia, 2021; Chopra et al., 2024), in different economic games (Kovarik, 2009; Kölle and Lauer, 2024; Kölle and Wenner, 2023) and proposed several theoretical

approaches (see Frederick et al. (2002) for a review). However, our focus is different compared to this literature in two crucial aspects. First, we do not study intertemporal decisions in a trade-off between giving money to someone else or keeping money for your own, excluding any self-interest. Second, while these papers study delayed realizations that reach from a couple of days to a year, our time span outreaches those papers by several years up to a decade. In our study, short-term outcomes are realized within one year, while long-term outcomes are realized within four to ten years.

Our project is linked to several strands of literature. In the literature of social and cognitive psychology, mental imagery is classified as episodic future thinking, which is one form of thinking about the future by projecting the self into the future (Atance and O'Neill, 2001). It has been shown that episodic future thinking can increase far-sighted choices in eating behavior and consumption in adults (Schacter et al., 2017; Peters and Büchel, 2010) as well as in children (Alan and Ertac, 2018), that it can facilitate empathy and pro-social intentions (Gaesser and Schacter, 2014; Gaesser et al., 2015; Yi et al., 2016; Gaesser et al., 2020) and that it is related to risk perceptions (Monroe et al., 2017; Bø and Wolff, 2020; Bordalo et al., 2022). Furthermore, the ability of imagining future events declines with aging (Zavagnin et al., 2016; Schacter et al., 2018) and is positively related to life satisfaction, work productivity and mental health (Eubanks et al., 2024). In this paper, we test an intervention of mental imagery to enhance people's future-mindedness and therefore shift people's decisions from short-term to long-term outcomes. We also look at heterogeneous effects of mental imagery and how they relate to individual characteristics.

Mental imagery has also been used as an intervention in economic settings. Ashraf et al. (2021) used an imagery-based entrepreneurial training program to improve economic outcomes of Colombian would-be entrepreneurs. They demonstrate that the ability of mental simulations positively correlates with economic outcomes, and that a training program which includes mental imagery and the emotional content of economic activity is more successful than a traditional business skills training. Thinking about and visualizing potential future outcomes has also helped to improve economic outcomes by increasing labor supply and educational spending (Orkin et al., 2023; Bernard et al., 2023). In organizations, mental imagery has been used to motivate employees to create vivid images of the future by triggering the experience-based cognitive system instead of using a meaning-based approach (Carton and Lucas, 2018). We implement mental imagery as an intervention to foster real-world support for long-term investments. These long-term investments mimic an intra- and interorganizational setting in which decision makers have to prioritize between short-term or long-term outcomes.

Another strand of literature deals with the mechanisms of how thinking about the future influences our current decision-making. Thinking about oneself in the future or engaging with a constructed version of one's future self has been used to achieve positive long-term outcomes in financial decisions (Hershfield et al., 2011), ethical decisions (van Gelder et al., 2013) and the health domain (Rutchick et al., 2018). The interventions aim to strengthen the future-self continuity by reducing the psychological distance to the future self (Hershfield, 2019). Similarly,

as mentioned above, mental imagery creates more concrete visualizations of future outcomes and therefore, reduces the psychological distance that is generated by the long time horizon (Trope and Liberman, 2010; Liberman and Trope, 2003). We test this mechanism to shed light on how mental imagery can shift people's decision preferences towards the long-term.

Finally, we contribute to the literature of how to implement different decision-making mindsets, and mental imagery in particular. For the implementation of mental imagery, many studies in social and cognitive psychology ask subjects to visualize themselves or specific situations, events and outcomes in the future (e.g. Peters and Büchel (2010); Gaesser et al. (2015, 2018)). Some studies rely on workshops, training programs or videos with which mental simulation and the visualization of future events are trained (Ashraf et al., 2021; Orkin et al., 2023; Bernard et al., 2023; Alan and Ertac, 2018). Others like Eubanks et al. (2024) use the Pragmatic Prospection Scale (Ruscio et al., 2023) which is a self-reported measure of future-focused thinking to assess the individual ability of pragmatic prospection. In this paper, we test a simple reading intervention to induce an imaginative mindset by making subjects think more concretely about potential outcomes. Importantly, we use standard instructions, that have already been used in the literature to successfully induce an intuitive and deliberative mindset (Barrafrem and Hausfeld, 2020; Bieleke et al., 2017) and make certain adjustments while keeping the format identical to induce an imaginative mindset.

The remainder of this paper is organized as follows: Section 2.2 explains the experimental design and the implemented treatments. Section 2.3 introduces a conceptual framework of how mental imagery fosters future-mindedness and states our hypotheses. Section 2.4 shows the results of our experiments, and Section 2.5 concludes.

## 2.2 Experimental Design

In order to study how preferences for short-term investments can be shifted towards long-term investments with minimal chances for one's future self to benefit from it directly, we conduct a lab experiment in which subjects can decide between investing a fixed amount of money in a short-term or a long-term project of the same sector. The subjects in our experiment are students and are presented with nine pairs of mostly real projects from the university of their current studies. [2] The projects were selected in collaboration with the university foundation. These reflect actual projects in the cultural, social, educational or technical sector that were either already existing or to be implemented within the next four to ten years in accordance with the university's long-term strategic vision. In each decision, subjects have to decide whether they want to invest the money in a short-term university project with a time horizon of one year or a long-term university project with a time horizon of four to ten years, depending on the project. Importantly, we made clear that the investment of the short-term project is realized

[2]Five decisions contained pairs of real university projects, while four decisions contained hypothetical university projects in order to vary the individual attractiveness of each project. The subjects did not know which projects are which. In that sense, all these projects were perceived as relevant for the organization they belong to and their related decision-making.

within one year. In contrast, the long-term projects are only realized after several years, such that the investment does not provide benefits to the decision-makers themselves. The correct execution of the short- and long-term payments are ensured by the financial department of the university.

We implement the problem of short-termism by offering a decision-making problem between benefits in the near and far future. All participants could in principle benefit of the short-term project of their organization, as the benefits are realized within their life cycle at their current university. In contrast, the benefits of the long-term projects are realized beyond the student's study time, serving the future generations of students. Subjects are not allowed to keep the money for themselves. The set-up mimics the situation of corporate management in an organizational context in which managers or board members do not have an outside option to benefit from their corporate budget directly, but rather have to decide between short-term and long-term investments within their organization. In addition, the incentive structure makes it more appealing to invest in the short-term projects, as the decision-makers themselves can benefit from the realization directly. An exemplary decision screen of the lab experiment is shown in Figure 2.1.

The projects cover a variety of domains, while the domain is kept constant within each pair of short- and long-term project. The full list of projects can be found in Table A2.1. The project pairs are displayed in random order and in the end one decision is randomly determined for each subject to be realized with an investment of 5 Euros.

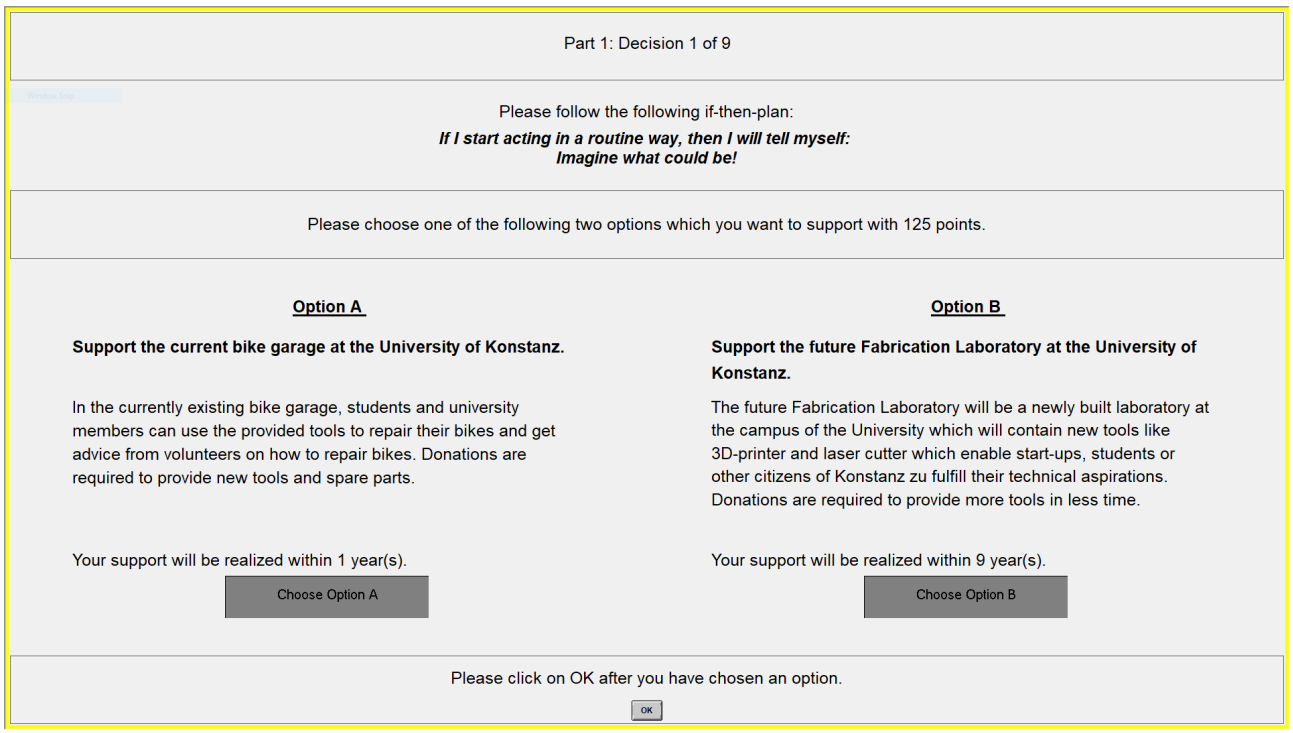


Figure 2.1: Decision Screen in the Imaginative Treatment

### 2.2.1 Treatments

In our experiment we interact four types of prospection-guidance interventions (intuitive, deliberate, imaginative, and control) with two decision-making settings (individual and group) in a 4 × 2 factorial design. We use a between-subject design to compare the effectiveness of our framework interventions and a within-subject design to study the differences of the framework interventions in individual and group decision-making. This means that within each framework variation, all subjects carry out the main task – nine investment decisions – first individually and then in a group setting.

#### Frameworks

In line with the three categories of prospection guidance proposed by Seligman et al. (2016) we manipulate the mindsets in which our subjects make their short- and long-term investment decisions. We use mental imagery to induce an imaginative mindset. Further treatments include an intuitive or deliberative mindset along with a control treatment in which no specific mindset is induced. Subjects in the control treatment had to draw geometric forms. The frameworks are induced via a standard reading intervention which consists of a short description of the decision-making guidance that subjects should follow during the experiment and which is implemented using the implementation intention plans by Gollwitzer (1999). We follow the procedure of previous studies who have successfully induced an intuitive and deliberative mindset by showing guided instructions at the beginning of an experimental task (e.g., Barrafrem and Hausfeld (2020); Bieleke et al. (2017)) and we construct the mental imagery treatment to follow the very same standard structure. We remind subjects of the framework on each decision screen during the course of the investment task of the experiment. An English translation of the framework used in our experiment can be found in Figure 2.2, where stacked phrases from top to bottom are adjusted for the imaginative (blue) / intuitive (red) / deliberative (yellow) treatment, respectively.

#### Decision-Making Setting

For each framework, subjects complete the investment decisions first individually, and then decide on the same decisions in fixed groups of three. The purpose of this treatment variation is to study a group investment task in board decision-making settings in an organizational context. The procedure of the group investment task is depicted in Figure 2.3. For each decision, first, each group member is informed about the current investment decisions. Next, they can chat for 75 sec to mimic the discussions of the board members. Afterwards, each group member has to decide between the short-term and the long-term option. The group decision is successful if it is unanimous, while groups have up to five trials per decision to reach unanimity. Otherwise, the group decision is not valid and no money will be invested in case the decision is randomly selected at the end of the experiment. Requiring unanimity for a successful group investment should mimic real-world budget conversations in organizations, in which a lack of

"In this experiment, we are especially interested in the effect of imagination / intuitive actions / cautious, deliberative actions on decisions.

Please avoid acting in a routine way / thinking too long / making decisions hastily. Instead, use your imagination / trust your instinct and listen to your gut feeling / carefully weigh the options and use your judgment in your decisions.

In order to make better decisions, it is helpful to use imagination / listen to one's gut feeling / use your mind.

According to researchers, it is helpful to make so called if-then-plans. Please follow the following if-then-plan:

As soon as I start to behave in a routine way / be insecure / act hastily, I will tell myself: Imagine what could be / Listen to your guts / Use your brain !

Please look at this plan and internalize it. To do so, please read out the plan three times in thought. Afterwards, please write down the plan on the provided sheet of paper."

Figure 2.2: Experimental instructions for mindset manipulation.

Note: Stacked phrases from top to bottom are adjusted for the imaginative, intuitive, deliberative treatment respectively. Colored boxes are used for visualization purposes in this figure only.

compromising skills can lead to suboptimal budget allocation. This procedure is repeated for each investment decision. Similarly to the individual decision tasks, the respective framework is repeated on each decision screen.

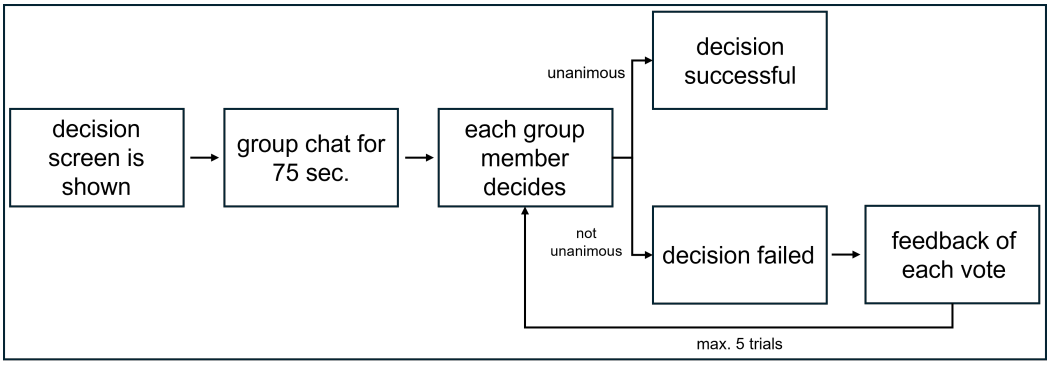


Figure 2.3: Chronological ordering of each group investment decision.

### 2.2.2 Remaining Tasks

After the main task – investment decisions individually and in groups – we elicit individual characteristics like time, risk and social preferences in an incentivized way as outlined below. As

a robustness check we ask for the unincentivized self-reported preferences respectively, following Falk et al. (2018) and the European Social Survey (Jowell et al., 2007). Table A2.3 in the Appendix lists a translated version of all self-reported measures used in the experiment.

In the time preference task we ask subjects to indicate in five decisions whether they want to receive a fixed payout next week or a higher payout in six months where we gradually increase the future payout over five decisions (Andersen et al., 2008). Afterwards, we ask subjects to subjectively state how impatient they are and whether they make future plans in general (Falk et al., 2018). In the risk preference task we ask subjects to select one out of nine games in which a die roll determines their payout. We gradually vary the riskiness and amount of the payout over nine games (Gneezy and Potters, 1997). In addition, subjects indicated how risk-averse / risk-seeking they see themselves (Falk et al., 2018). Finally, we elicit subjects' Social Value Orientation (SVO) over six decisions (Murphy et al., 2011) and asked them afterwards about their attitude in altruism, reciprocity and trust (Falk et al., 2018). Subjects were informed that their role in the SVO task will be randomly determined and that they will be matched with another subject. At the end, we run a short questionnaire, where we cover BIG-5 personality traits (Gosling et al., 2003), future optimism, judgments of the frameworks and projects, and socio-demographic characteristics. Subjects were paid according to randomly determined decisions in the time, risk and social preference tasks and received an additional show-up fee.

### 2.2.3 Procedure

The experiment was programmed in "zTree" (Fischbacher, 2007) and run at the computer laboratory "Lakelab" of the University of Konstanz. In total, 240 students (mean age: 23, 52.5% female) were recruited with the software "hroot" (Bock et al., 2014) such that 60 students participated in each treatment (imaginative, intuitive, deliberative, control). The sessions were conducted between November 2021 and June 2022. [3]

### 2.2.4 Online Experiment

In addition, we conduct an online experiment with a representative sample of the German working population ($N = 489$, mean age: 44.5, 48.5% female) for several purposes. First, we want to extend our results to a different context. The sample consists of employed subjects from the German working population, and it is chosen according to representative quotas on gender, age, region (West Germany / East Germany) and education. Table A2.4 in the Appendix shows the composition of the online sample on a variety of characteristics, including employment status, industry and leadership responsibilities.

Next, the set-up of the online experiment moves away from the organizational context in which the decision-makers make investment decisions about their own organization, and towards

[3]Because of Covid-19 restrictions, the sessions couldn't be finished in winter 2021 and had to be resumed in summer 2022.

a context in which decision-makers are less involved and decide on investments across several organizations. Participants in the online experiment are presented with seven real investment project pairs in Germany, each again consisting of a short-term project with a time horizon of one year and a long-term project with a time horizon of four to ten years. Each short- and long-term project of a decision comes from the same organization and the organizations cover cultural, environmental, social, sports, technical and educational domains. Importantly, in contrast to the lab experiment, these investment projects do not come from a single organization to which all the subjects belong. These projects come from a variety of organizations across the whole country. Table A2.2 in the Appendix lists all organizations and the corresponding projects that were selected for the online experiment.

Finally, we test specific treatment variations to get a better picture of the effect of mental imagery. Subjects in the online experiment received the same implementation intention plans as in the lab experiment. We changed the control treatment into a pure informational treatment without any guidance and added another mental imagery treatment to test a variation in intensity of mental imagery. In addition, subjects were asked for their long-term preference in one project pair in case the time horizon of the long-term project was changed. Subjects did not play in groups. Time, risk and social preferences were elicited in a non-incentivized way, and we collected socio-demographic characteristics. In the end, ten subjects were randomly selected and one investment decision of 20 Euros was carried out for each subject. Table 2.1 summarizes the similarities and differences between the lab and the online experiment.

Table 2.1: Design Comparison Lab and Online Experiment

| | Lab | Online |
|---|---|---|
| Sample | University students | General German working population |
| Projects' organization | University of Konstanz | Different organizations across Germany |
| Treatments | Imaginative<br>Intuitive<br>Deliberative<br>Control | Imaginative (Base)<br>Imaginative (Vivid)<br>Intuitive<br>Deliberative<br>Control |
| Implementation of Mental Imagery (Base) | "As soon as I start to behave in a routine way, I will tell myself Imagine what could be!" | |
| Implementation of Mental Imagery (Vivid) | | "$\cdots$ I will tell myself: Imagine what could be! Close your eyes and create a concrete image – how does it look like and how does it feel?" |
| Implementation of Control Treatment | Drawing geometric figures | Pure informational instructions |
| Group Task | ✓ | × |

The online experiment was programmed in "Qualtrics" and run with the help of the survey company "Bilendi". Data collection was carried out in December 2022 and subjects took 11.3 minutes on average and were compensated by the survey company. The survey company initially collected responses of 540 subjects, as they offer an additional 10% of collected observations for quality issues. We exclude one subject due to missing data. In addition, we exclude the fastest and slowest 5% quantile of subjects in the main task in each treatment to match the suggested quality rules of the survey company and follow the procedure of previous literature (e.g., Bellani et al. (2021); Lobeck and Støstad (2023)). We are left with 489 deliberately unequally but randomly distributed across five treatments to focus on the effect of mental imagery (Control = 86, Deliberative = 84, Intuitive = 89, Imaginative Base = 160, Imaginative Vivid = 70).

## 2.3 Conceptual Framework

In this section, we present a conceptualization of the effect of mental imagery on long-term choices by adapting the insights of Construal Level Theory (CLT) by Trope and Liberman (2010) to our setting. The essence of CLT is that people form mental construals of objects, that they cannot experience in the here and now. These objects include past and future events, other places or people, and counterfactuals of reality. Mental construals are distinct from direct experience and allow people to transcend the immediate situation to capture psychologically distant objects. Furthermore, mental construals vary in their level of abstractness and therefore define how abstract or concrete the representation of an object is. Low-level construals contain more detailed information and are more concrete representations, whereas high-level construals concentrate on central features and are more abstract representations.

The level of construal is influenced by the psychological distance towards the object. The psychological distance is composed of the temporal, spatial, social and hypothetical distance. Temporal distance therefore influences the level of construals. People create high-level construals for objects in the far future and low-level construals for objects in the near future (Liberman and Trope, 2003).

**Conjecture 1** *The (temporal) distance influences how concrete or abstract the representation of the object is.*

Furthermore, the relationship between the distance towards an object and the corresponding level of construal is bilateral. Not only does the distance to an object influence the level of construal, but the reverse is true as well. As Trope and Liberman (2010) state on p. 442: "$\cdots$ different levels of construal serve to expand and contract one's mental horizons and thus mentally traverse psychological distances." This means that people infer the distance to an object by their mental representation of the object. Concrete representations of objects let people interpret these objects as close, while abstract construals are interpreted as being distant.

**Conjecture 2** *The level of construal influences the perceived (temporal) distance.*

Moreover, the construal level of an object can not only be influenced by the distance towards the object, but also by the form of representation. The authors claim that a mental representation based on words induce more abstract and therefore high-level construals, while a mental representation based on pictures generates more concrete and therefore low-level construals. Mental imagery is the act of creating concrete images of objects in the mind's eye by imagining specific details of the outcome (Nanay, 2021). Therefore, mental imagery lowers the level of construal by generating concrete representations of the outcome.

**Conjecture 3** *Mental imagery can change the level of construal.*

Taking the implications of the three conjectures of CLT together, we expect the following mechanism in our experiment which is also depicted in Figure 2.4: First, long-term projects have a higher temporal distance and are therefore represented in more abstract construals compared to short-term projects (Conjecture 1). Next, an intervention of mental imagery creates more concrete representations of the long-term project, which lowers the construal level. (Conjecture 3). Due to the lower construal level, the psychological distance towards the long-term project is reduced (Conjecture 2). This means that subjects interpret the lower construal level of the long-term project as indicating that the object is less distant. Choices of subjects are influenced by the perceived distance towards the projects. In line with the literature review on short-termism in Section 2.1, subjects prefer to take investment decisions for projects with a lower distance compared to projects with a higher distance. In summary, under mental imagery, people perceive the long-term option as more concrete, therefore less distant, which in turn leads to more long-term choices compared to the other treatments.

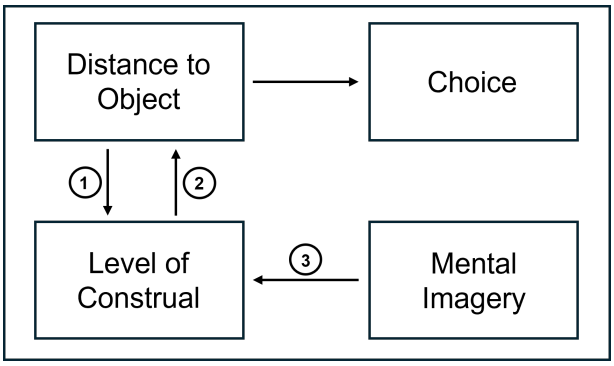


Figure 2.4: Schematic illustration of the conceptual framework

### 2.3.1 Hypotheses

Our main variable of interest is the choice for long-term projects. In line with our pre-analysis plan, we aggregate the nine decisions in the main task for each framework (separately for individual and collective decisions) by taking the share of long-term decisions (denoted $\bar{y}$ below) for hypothesis testing. In addition, we analyze the raw data with the help of regressions with robust clustered standard errors.

**Impact of mental imagery in individual decision-making (between-subject):**

**Hypothesis 1** *We expect more individual subjects to choose the long-term projects in the imaginative framework compared to the other three treatments. We will examine the differences among the other three frameworks in an exploratory manner (H1).*
$H_0 : \bar{y}^{imag.} = \bar{y}^{others}, H_1 : \bar{y}^{imag.} > \bar{y}^{others}$

We expect these effects to remain robust to controlling for a variety of individual-level and situation-specific control variables in regression analyses. We will also use the control variables to explain the heterogeneity in the effectiveness of our framework interventions.
Our first hypothesis is based on the findings of the positive influence of mental imagery on future-mindedness stated in Section 2.1 and the conceptualization described in this section.

**Impact of decision-making setting (within-subject):**

**Hypothesis 2** *We expect that long-term projects will be selected less often in groups compared to individuals (H2A).*
$H_0 : \bar{y}_{indiv.} = \bar{y}_{group}, H_1 : \bar{y}_{indiv.} < \bar{y}_{group}$
*We expect this difference to be less pronounced in the imaginative framework (H2B).*
$H_0 : \bar{y}^{imag.}_{indiv.} = \bar{y}^{imag.}_{group}, H_1 : \bar{y}^{imag.}_{indiv.} < \bar{y}^{imag.}_{group}$

We also expect these effects to remain robust to controlling for a variety of control variables in regression analyses. We will aggregate the control variables at a group level and use them to explain the heterogeneity in the differences between individual and group decisions.
Our hypotheses H2A and H2B are based on the vast literature on group decision-making biases (median shift, groupthink, polarization, emotional contagion; see e.g. Bénabou (2013), for a review) which could hinder prospection – and the beneficial effects of imaginative prospection in particular.

## 2.4 Results

We want to study how long-term decisions can be influenced by testing different framing interventions, particulary mental imagery. We structure our analysis by performing a manipulation check to assess if mindsets have successfully been adopted (Section 2.4.1). We continue by analyzing the effect of mental imagery on long-term choices (Section 2.4.2). We shed light on potential mechanisms (Section 2.4.3) and study the heterogeneous effects of mental imagery (Section 2.4.4). Finally, we look at the influence of mental imagery on long-term decisions in groups (Section 2.4.5) and provide some robustness checks (Section 2.4.6).

### 2.4.1 Framework Manipulation Check

First, we check whether we successfully put subjects into the specific mindsets by looking at the response times. Figure 2.5 shows the average response time that subjects in the lab experiment needed per individual decision across mindsets.

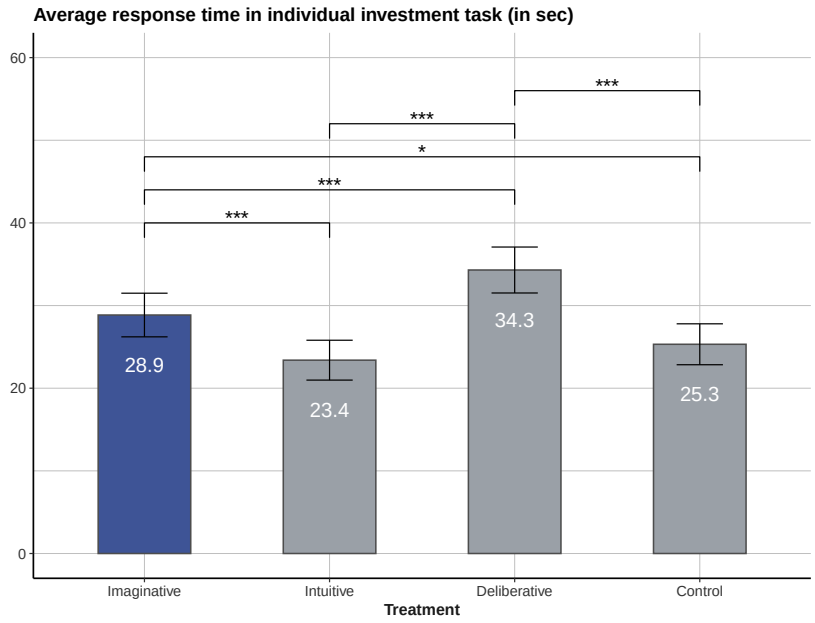


Figure 2.5: Average response times in individual investment task (in sec)

Note: Black whiskers represent 95% confidence intervals. Wilcox Rank-Sum Tests used for treatment comparison.

Subjects in the deliberative framework take 34.3 sec on average (median = 33.8 sec, SD = 11.0 sec) to decide between a short-term and a long-term project, which is significantly more time than in the other frameworks (all *p-values* $< 0.01$). [4] As subjects in the deliberative treatment are explicitly asked to take their time while deciding, this is in line with the intended manipulation. In the intuitive treatment, subjects take 23.4 sec per decision on average (median = 21.5 sec, SD = 9.5 sec) which is faster than subjects in a deliberative or imaginative mindset (*p-values* $< 0.01$). Our implementation of the imaginative mindset leads to an average response time of 28.9 sec (median = 27.3 sec, SD = 10.4 sec) which is in between the deliberative and the intuitive mindset. The response time analysis confirms that subjects successfully adopted the respective mindset.

**Result 1** *Subjects in the intuitive treatment take less time, while subjects in the deliberative treatment take more time on average for each investment decision.*

### 2.4.2 Impact of Mental Imagery

Our main hypothesis is that an imaginative framework, induced via mental imagery, fosters future-mindedness and shifts people's choices to more long-term decisions. Therefore, we expect subjects in the imaginative treatment to decide more often in favor of the long-term project compared to the other treatments. Figure 2.6 shows the share of long-term decisions across the four treatments when subjects decided individually.

[4] If not indicated differently, the average over all nine decisions per subject is computed and non-parametric Wilcox Rank-Sum tests are used for hypothesis testing.

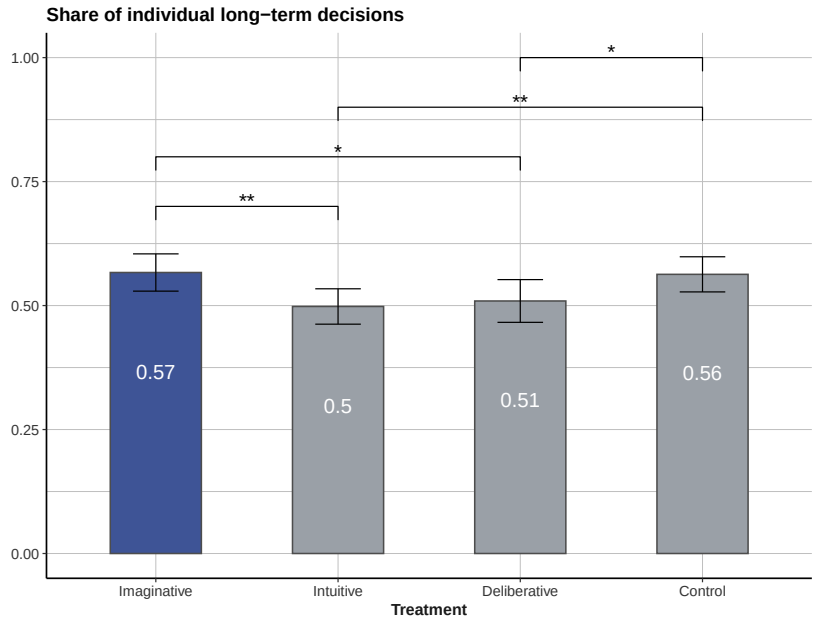


Figure 2.6: Share of long-term decisions in individual setting of lab experiment across treatments.

Note: Black whiskers represent 95% confidence intervals. Wilcox Rank-Sum Tests used for treatment comparison.

In an imaginative mindset, 57% of all decisions are taken in favor of the long-term option. This is significantly more compared to the intuitive and deliberative mindset (50% and 51%, *p-values* $< 0.06$). However, the imaginative mindset does not lead to more long-term decisions compared to the control treatment in which subjects had to draw geometric forms (56%, $p = 0.889$). Therefore, we find partial evidence for hypothesis H1. In fact, in the intuitive and deliberative treatment, the long-term option is also taken significantly less often compared to the control treatment (*p-values* $< 0.07$). The results show that an intuitive and a deliberative mindset lead to a reduction in long-term preference, while the induction of an imaginative mindset does not lower the preference for the long-term compared to a control treatment. We discuss potential explanations and design variations in Section 2.4.6.

Our results can also be seen in Model (1) and (2) in regression Table 2.2 and hold when including a variety of covariates to control for subject- and project-specific characteristics. [5] The table shows linear probability models to explain the probability of choosing the long-term project in comparison to the control treatment, which serves as the baseline.[6] For each model, the second column depicts the p-values when comparing the respective treatment coefficient to the coefficient of the imaginative treatment via linear hypothesis tests. All results are shown with robust standard errors clustered at the subject level. Model (1) and (2) show that the

[5]While subjects favor long-term projects in the cultural domain more than in the technical or educational domain, there is no systematic treatment variation. The complete regression table including a full list of all covariates can be found in Table A2.5 in the Appendix.

[6]Note that linear probability models are used despite binary outcome variables for simpler interpretation. All regression models with binary dependent variables have been tested with probit models, and the results stay qualitatively the same.

likelihood of choosing the long-term option is lower in the deliberative and intuitive treatment compared to the control treatment and the imaginative treatment. Mental imagery leads to an increase of 7% in the probability of choosing the long-term option compared to an intuitive or deliberative mindset.

Table 2.2: Treatment Effect on Long-Term Choice in Lab Experiment (Linear Probability Models)

| | Dependent Variable: Choose Long-term Project | | | |
|---|---|---|---|---|
| | Model (1) | | Model (2) | |
| | Coef. | vs. Imag. | Coef. | vs. Imag. |
| Intuitive | -0.065** | p = 0.024 | -0.057** | p = 0.032 |
| | (0.026) | | (0.028) | |
| Deliberative | -0.054* | p = 0.059 | -0.054* | p = 0.038 |
| | (0.028) | | (0.029) | |
| Imaginative | 0.004 | | 0.010 | |
| | (0.026) | | (0.027) | |
| Constant | 0.563*** | | 0.485*** | |
| | (0.018) | | (0.109) | |
| Covariates | No | | Yes | |
| Observations | 2,160 | | 2,097 | |
| $R^2$ | 0.004 | | 0.050 | |
| Adjusted $R^2$ | 0.002 | | 0.039 | |

Note: Linear probability models with binary long-term choice as dependent variable. The regressors are indicators for the intuitive, deliberative and imaginative treatment respectively. Covariates include subject- and project-specific characteristics (a full list can be found in Table A2.5 in the Appendix). The second column of each model shows the p-values of linear hypothesis tests when comparing the respective coefficient to the coefficient of the imaginative treatment. Standard errors are clustered at the individual level.
*p<0.1; **p<0.05; ***p<0.01

Our results suggest that an image-driven way of thinking and deciding about the future can increase the propensity of long-term decisions compared to other established mindsets such as an intuitive or deliberative mindset. Remarkably, a deliberative mindset, which can arguably be seen as the predominant environment in which companies and organizations operate, consistently leads to less long-term decisions.

**Result 2** *Mental imagery leads to more long-term choices compared to an intuitive or deliberative treatment.*

### 2.4.3 Mechanism

To find out how mental imagery can lead to more long-term decisions compared to other mindsets, we want to study different channels through which future-mindedness might be affected, including the proposed mechanism stated in Section 2.3. At the end of the experiment, we elicited and collected different individual preferences, traits, and characteristics (see Section

2.2.2). These characteristics help us in two ways: First, we can investigate which traits in general influence long-term decisions and second, which subjects are especially influenced by mental imagery. Table A2.6 and Table A2.7 in the Appendix show the influence of these individual characteristics (time, risk, social preferences, future orientation and optimism, personality traits and socio-demographics) on the probability of choosing the long-term option in linear probability models. Table 2.3 shows a selection of these regression tables.

Table 2.3: Influence of individual characteristics on long-term choice

| | Dependent Variable: Choose Long-term Project | | | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| Risk-Seekingness | −0.008** | | | | |
| | (0.004) | | | | |
| Trust | | 0.013** | | | |
| | | (0.006) | | | |
| Future Optimism | | | 0.022*** | | |
| | | | (0.007) | | |
| Involvement LT Project | | | | 0.076*** | |
| | | | | (0.016) | |
| Excitement LT Project | | | | | 0.113*** |
| | | | | | (0.017) |
| Constant | 0.567*** | 0.475*** | 0.523*** | 0.508*** | 0.299*** |
| | (0.019) | (0.029) | (0.010) | (0.014) | (0.040) |
| Observations | 2,160 | 2,160 | 2,160 | 1,404 | 648 |
| $R^2$ | 0.002 | 0.002 | 0.003 | 0.011 | 0.066 |
| Adjusted $R^2$ | 0.001 | 0.001 | 0.002 | 0.011 | 0.065 |

Note: Linear probability models with binary long-term choice as dependent variable. *Risk-Seekingness* is an incentivized measure of risk preferences. The remaining independent variables are self-reported measures, and the elicitation method can be found in Table A2.3 in the Appendix. Standard errors are clustered at the individual level. *p<0.1; **p<0.05; ***p<0.01

The characteristics that clearly influence the long-term choice are risk preferences, social preferences (trust in people), future optimism and the involvement in and the excitement for the respective long-term project. For all mentioned covariates, except for the risk preference, the effect is as expected. Higher trust, future affinity, and project involvement and excitement increase the probability of a long-term choice, irrespective of the treatment. Surprisingly, more risk-averse subjects are more likely to choose the long-term option. Due to the random assignment of subjects into treatments, these individual characteristics should not differ across treatments and therefore, cannot explain why mental imagery leads to more long-term choices than an intuitive or a deliberative mindset. In addition, we test if there are treatment differences for the stated covariates, as one might assume that some characteristics (e.g., future optimism) might be influenced by the induced mindsets and that the effect varies across treatments. Table A2.8 in the Appendix shows that the characteristics are not affected by the treatments and therefore, we do not find any systematic variation or effect on individual characteristics that could explain the main treatment effect.

**Result 3** *Subjects with a higher trust in people, future optimism, risk aversion and project involvement and excitement are more likely to choose the long-term project.*

Having shown that our treatments did not change any time or future related preference, we now turn to the conceptualization described in Section 2.3. Even if the individual time preferences have not changed, we suggest that the perception of the time horizon of the long-term projects are differently affected by the treatments, as the decision is not about their own payoff. According to our model, we expect the imaginative mindset to create more concrete construals of the long-term projects and therefore reduce the perceived temporal distance of the long-term project. Thus, subjects are more likely to choose the long-term option compared to the other treatments. In our experiment, the time horizon of the long-term option varied between four and ten years. Figure 2.7 shows the predicted probability of choosing the long-term option for each treatment, depending on the time horizon of the long-term project.

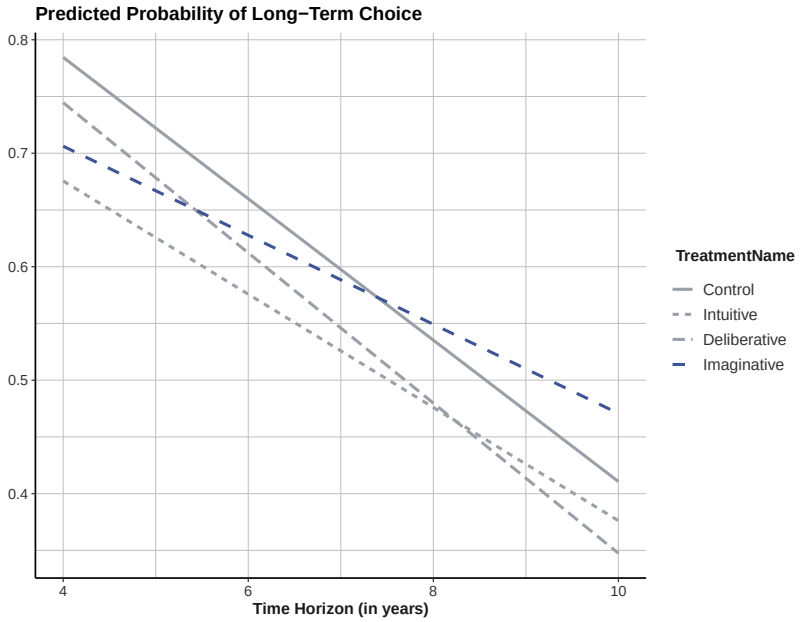


Figure 2.7: Predicted probability of choosing long-term project dependent on time horizon across treatments in the lab.

Note: The graph is based on the linear probability model: $LT\ Choice_i = \beta_0 + \beta_1 * Treatment_j + \beta_2 * Time_i^{LT} + \beta_3 * Treatment_j * Time_i^{LT} + \epsilon_i$, where the dependent variable indicates a binary long-term choice, $Treatment_j$ indicates the respective treatment and $Time_i^{LT}$ denotes the time horizon of the long-term project. Standard errors are clustered at the individual level.

The predicted values of choosing the long-term option ($LT\ Choice$) in each decision $i$ are based on the following linear regression model in which $Treatment_j$ indicates the respective treatment $j \in$ (imaginative, intuitive, deliberative) and $Time_i^{LT}$ is the time horizon of the long-term option in the respective decision $i$: [7]

$$LT\ Choice_i = \beta_0 + \beta_1 * Treatment_j + \beta_2 * Time_i^{LT} + \beta_3 * Treatment_j * Time_i^{LT} + \epsilon_i \qquad (2.1)$$

[7] We checked the results allowing for a non-linear effect of the time horizon and the results stay qualitatively the same.

Two insights can be taken from the graph. First, the longer the time horizon, the less likely it is chosen, irrespective of the treatment. This supports our assumption that subjects generally prefer projects with a lower temporal distance. However, the graph also shows, that there is a treatment effect in the perception of the time horizon, although the effect is not significant. In the imaginative treatment, long-term projects with a particular long time horizon tend to be chosen more often compared to other treatments.

To explore the potential mechanism more thoroughly, we analyze how the share of long-term decisions depends on the time horizon of the long-term project. Figure 2.8 shows how often the long-term option is chosen on average for long-term project with a rather short time horizon (four to seven years) compared to long-term project with a large time horizon (eight to ten years). Figure 2.8A includes all decisions in which the time horizon of the long-term option is below the median time horizon, and Figure 2.8B reports the results in which the time horizon of the long-term option is above the median time horizon.

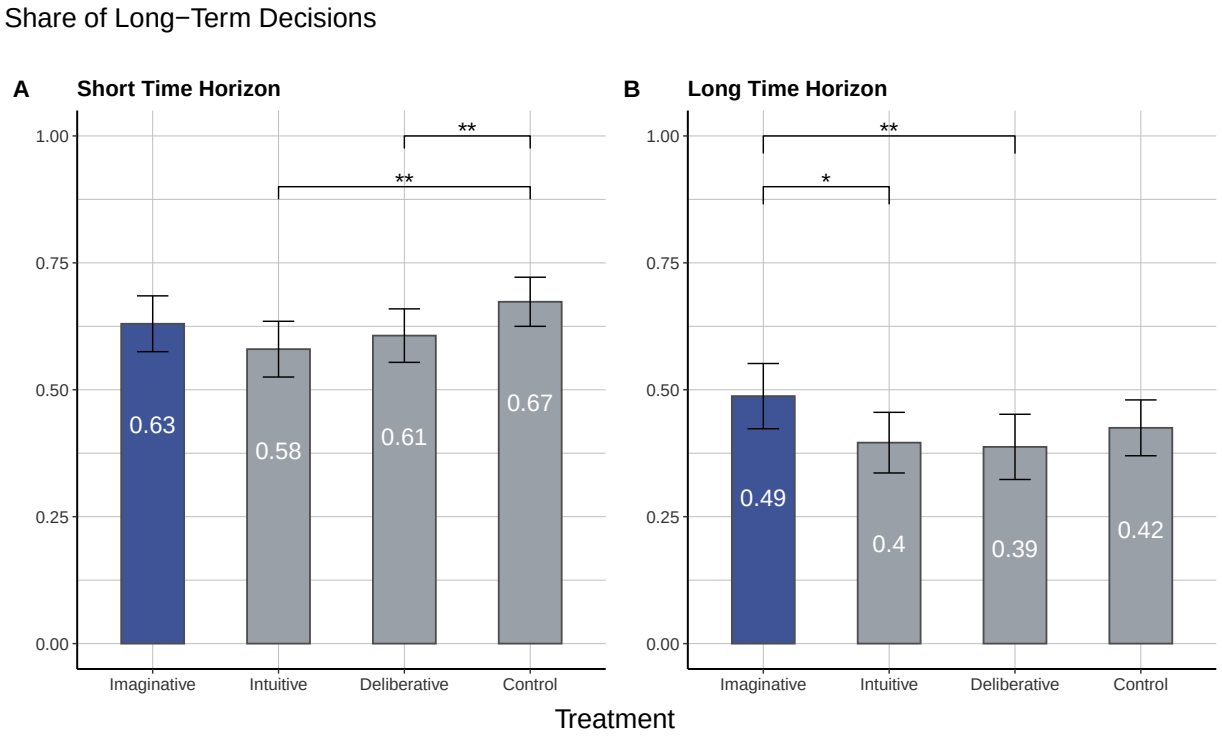


Figure 2.8: Share of long-term decisions across treatments. Panel A shows decisions in which long-term projects have a short time horizon (4 - 7 yrs.), while Panel B shows decisions in which long-term projects have a long time horizon (8 - 10 yrs.)

Note: Black whiskers represent 95% confidence intervals. Wilcox Rank-Sum Tests used for treatment comparison.

As already mentioned, across all treatments, the share of long-term decisions for projects with a time horizon above the median is significantly lower compared to projects with a time horizon below the median. Projects with a higher time horizon are chosen less often. However, there are treatment differences. For decisions which have shorter time horizons, an intuitive and deliberative mindset lead to fewer long-term decisions compared to the control treatment

(see Figure 2.8A). By contrast, in the imaginative treatment, subjects choose the long-term option significantly more often compared to other mindset when the time horizon is especially large (see Figure 2.8B). Although there is no difference in long-term preference between the imaginative and the control treatment when considering all decisions, there is a clear difference in the mechanism of both treatments. Without any mindset, subjects show a higher preference for long-term projects with a shorter time horizon compared to an intuitive or deliberative mindset. When the time horizon of the long-term project is especially large, an imaginative mindset leads to more long-term decisions compared to the other mindsets.

This result can also be seen in Table 2.4 showing the regression output of a linear probabtility model. Long-term projects that have a particularly high time horizon are almost 25% less likely to be selected in the control treatment.

Table 2.4: Influence of time horizon on long-term choice and treatment interaction

| | *Dependent variable:* | |
|---|---|---|
| | Choose Long-Term Project | |
| | (1) | (2) |
| Intuitive | $-0.093^{**}$ | $-0.091^{**}$ |
| | (0.037) | (0.039) |
| Deliberative | $-0.067^{*}$ | $-0.062$ |
| | (0.036) | (0.038) |
| Imaginative | $-0.043$ | $-0.038$ |
| | (0.037) | (0.039) |
| Long Time Horizon | $-0.248^{***}$ | $-0.699^{***}$ |
| | (0.038) | (0.071) |
| Intuitive x Long Time Horizon | 0.064 | 0.077 |
| | (0.059) | (0.061) |
| Deliberative x Long Time Horizon | 0.029 | 0.016 |
| | (0.055) | (0.057) |
| Imaginative x Long Time Horizon | $0.106^{*}$ | $0.108^{*}$ |
| | (0.060) | (0.062) |
| Constant | $0.673^{***}$ | $0.890^{***}$ |
| | (0.025) | (0.116) |
| Covariates | No | Yes |
| Observations | 2,160 | 2,097 |
| $R^2$ | 0.044 | 0.103 |
| Adjusted $R^2$ | 0.041 | 0.091 |

Note: Linear probability models with binary long-term choice as dependent variable. The regressors are indicators for the intuitive, deliberative and imaginative treatment respectively. *Long Time Horizon* is a dummy variable, indicating if the time horizon of the long-term project is above the median. Covariates include subject- and project-specific characteristics (a full list can be found in Table A2.5 in the Appendix). Standard errors are clustered at the individual level. $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01

However, in the imaginative treatment, this negative effect is partly mitigated by 11% (see Model (1) in Table 2.4) meaning that projects with a large time horizon are chosen only 14% less in the imaginative treatment. This result is robust to including a variety of individual- and project-specific controls (see Model (2) in Table 2.4).

Taken together, mental imagery can create more concrete construals of projects with a large time horizon, and therefore influences the perception of the time horizon of the long-term options. These low-level construals reduce the perceived temporal distance of the project, leading to more long-term choices when the time horizon is particularly long.

**Result 4** *Mental imagery leads to more long-term decisions for projects with a particularly long time horizon.*

### 2.4.4 Heterogeneous Effects of Mental Imagery

After having analyzed the main mechanism through which mental imagery leads to more long-term decisions, we want to shed light on individual heterogeneity of this mechanism. The question we are interested in is who is affected by mental imagery. At the beginning of Section 2.4.3 we discussed that future optimism positively influences the probability of choosing the long-term option. We asked subjects in our post experimental questionnaire about their present and future (in 10 years) life satisfaction rating in general. As per our definition, subjects that rate their future life satisfaction higher than their current one are called future optimists, while subjects that rate these two components equally or even in the opposite direction are called future realists / pessimists (henceforth future pessimists). In our experiment, 45% of subjects are categorized as future optimists, which is similarly distributed across treatments. Over all treatments, future optimists are more likely to choose the long-term option (see Table A2.6 in the Appendix). Figure 2.9 shows how the predicted probability of choosing the long-term option depends on the time horizon of the long-term option for future pessimists (Figure 2.9A) and future optimists (Figure 2.9B). Figure 2.9A illustrates that future pessimists are negatively influenced by the time horizon across all treatments. By contrast, future optimists that are treated with mental imagery show a less negative effect of the time horizon.

This effect is also captured in Table 2.5 which shows for each treatment how future optimists and future pessimists differ in their perception of the time horizon and how this in turn influences their long-term decisions. We can see that for future pessimists, the probability of choosing the long-term option sharply decreases if the time horizon of the project is large, irrespective of the treatment. This result also holds for future optimists, except for those in the mental imagery treatment. This means that future optimists that were treated to have an imaginative mindset are not negatively effected by a large time horizon of the long-term choice. [8]

**Result 5** *Mental imagery works especially well for subjects that are classified as future optimists.*

[8] Note that this is not due to a selection effect in different treatments, as we have seen in Table A2.8 in the Appendix that future optimism does not vary by treatments.

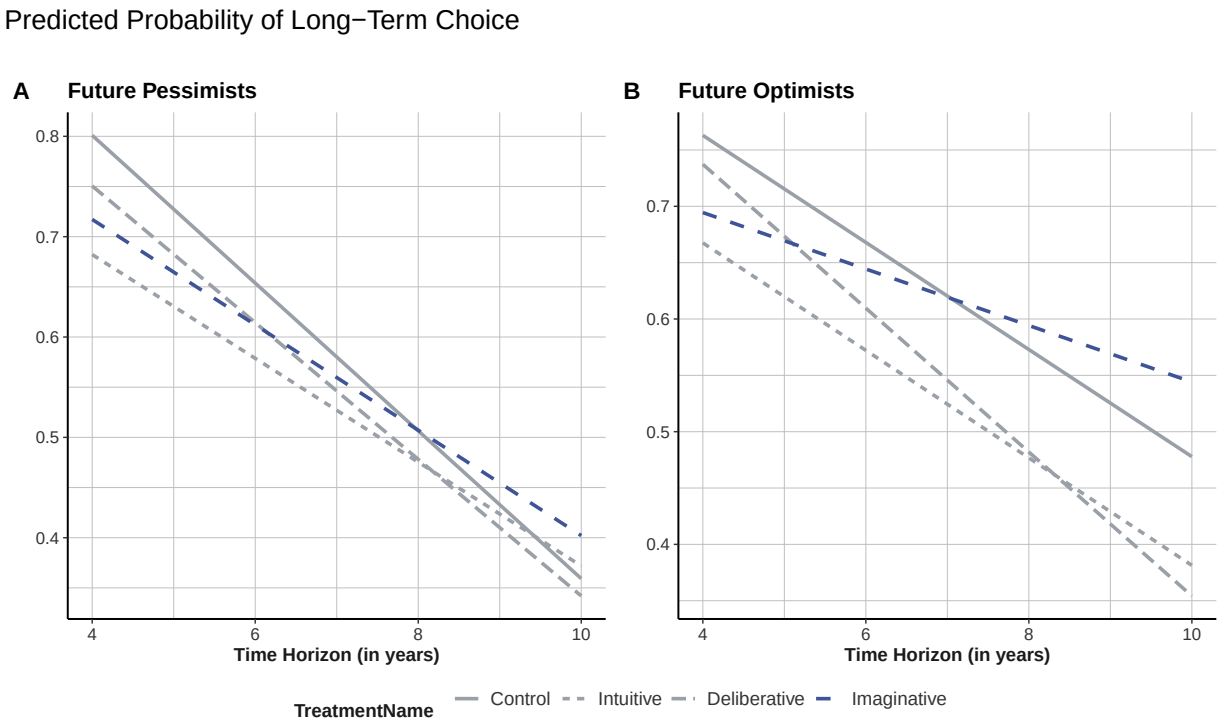


Figure 2.9: Predicted probability of choosing long-term project dependent on time horizon across treatments in the lab for future pessimists (Panel A) and future optimists (Panel B)

Note: The graph is based on the linear probability model: $LT\ Choice_i = \beta_0 + \beta_1 * Treatment_j + \beta_2 * Time_i^{LT} + \beta_3 * Treatment_j * Time_i^{LT} + \epsilon_i$, where the dependent variable indicates a binary long-term choice, $Treatment_j$ indicates the respective treatment and $Time_i^{LT}$ denotes the time horizon of the long-term project. The model is computed separately for future optimists and future pessimists. Future optimists are defined as people who value their future life satisfaction higher than their current life satisfaction. Standard errors are clustered at the individual level.

Table 2.5: Interaction of time horizon and future optimism

| | *Dependent variable:* | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Choose Long-term Project | | | | | | | |
| | Imaginative | | Intuitive | | Deliberative | | Control | |
| | Opt. | Pess. | Opt. | Pess. | Opt. | Pess. | Opt. | Pess. |
| Long Time Horizon | −0.066 | −0.215*** | −0.130* | −0.229*** | −0.169*** | −0.261*** | −0.187*** | −0.296*** |
| | (0.071) | (0.061) | (0.069) | (0.061) | (0.057) | (0.056) | (0.062) | (0.048) |
| Constant | 0.634*** | 0.626*** | 0.556*** | 0.600*** | 0.585*** | 0.624*** | 0.677*** | 0.671*** |
| | (0.046) | (0.034) | (0.046) | (0.035) | (0.037) | (0.039) | (0.027) | (0.039) |
| Observations | 261 | 279 | 243 | 297 | 243 | 297 | 234 | 306 |
| $R^2$ | 0.004 | 0.046 | 0.017 | 0.052 | 0.028 | 0.067 | 0.036 | 0.087 |
| Adjusted $R^2$ | 0.001 | 0.042 | 0.013 | 0.048 | 0.024 | 0.064 | 0.031 | 0.084 |

Note: Linear probability models with binary long-term choice as dependent variable. *Long Time Horizon* is a dummy variable, indicating if the time horizon of the long-term project is above the median. Regressions are computed separately for future optimists (*Opt.*) and future pessimists (*Pess.*) for each treatment. Future optimists are defined as people who value their future life satisfaction higher than their current life satisfaction. Standard errors are clustered at the individual level. *p<0.1; **p<0.05; ***p<0.01

### 2.4.5 Long-term Decisions in Groups

We now turn our attention to the effect of mental imagery on long-term decisions in groups. To assess the effect of a group decision-making setting, we compare actual group decisions with hypothetical group decisions. Hypothetical group decisions are defined as group decisions that would have resulted by majority rule if each group member had chosen according to their choice in the individual decision-making setting. Figure 2.10 shows the share of long-term decisions in groups for actual and hypothetical outcomes across treatment. Against our expected hypotheses H2A and H2B, a group decision-making setting has no influence on the long-term choice in any of the treatment.

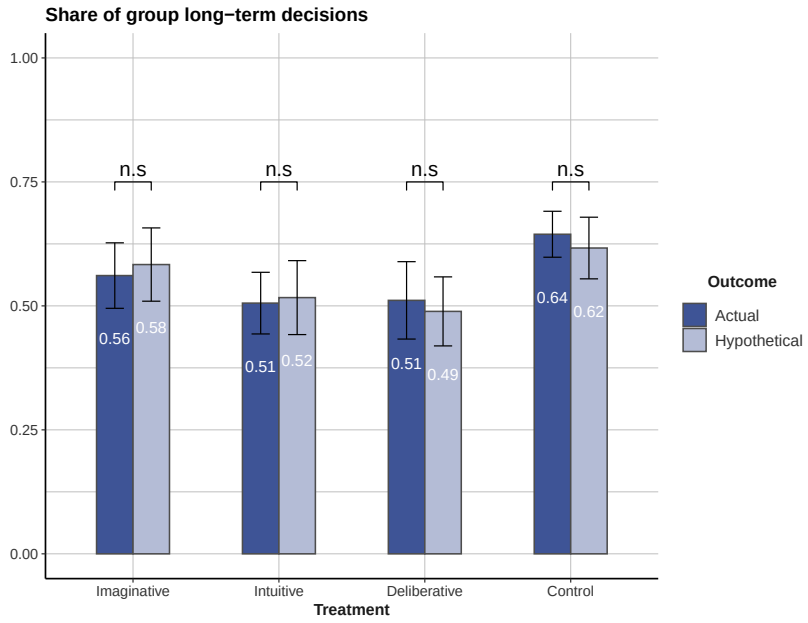


Figure 2.10: Effect of group decision-making setting on long-term choice.

Note: Black whiskers represent 95% confidence intervals. *Actual* shows the share of long-term decisions in groups. *Hypothetical* indicates the hypothetical share of long-term decisions if group members would decide according to their individual choice and the group outcome would be determined via majority rule. Wilcox Signed-Rank Tests used for hypothesis testing.

Furthermore, an imaginative mindset does not have a positive impact on collective long-term decisions compared to other mindsets. A potential explanation could be the lack of saliency of the frameworks in the group task, since subjects might be focused more on the repeating chat conversations with their group members. Mental imagery could be a very individual practice which could be easily overridden in group settings. Another explanation could be the structure of our experiment, since subjects first had to explicitly decide on each project pair individually before deciding in groups.

In the control treatment, 64% of group choices are in favor of the long-term choice, which is significantly more compared to group decisions in the other mindsets ($p\text{-}values < 0.06$). How-

ever, this is mainly due to a difference in average group composition and shows the importance of individual preferences for collective long-term decisions. We define the group composition as the number of group members that have a preference for the long-term choice indicated by their choice in the individual task at the beginning of the experiment. In the control treatment, there are more decisions in which two or more subjects individually chose the long-term option compared to the intuitive or deliberative treatment ($p\text{-}values < 0.05$) which then results in more collective long-term decisions.

To shed more light on the effect of group composition on collective long-term choices, Figure 2.11 shows the share of long-term decisions of groups pooled over all treatments, depending on the respective group composition. The share of long-term decisions in groups increases in the number of group members that have an individual preference for the long-term option. Of all group decisions, in which no group member has an individual preference for the long-term choice, only 4% end up in a long-term choice. This goes up to a share of 92% of all group decisions in which all group members initially have an individual long-term preference. The biggest increase in long-term choices is given for a group composition of two individuals who are in favor of the long-term option, compared to a group composition where only one member is in favor of the long-term option (77% compared to 28%). This means that as soon as a majority of the group has an individual preference for the long-term choice, the share of long-term decisions in groups increases sharply.

**Result 6** *The initial group composition influences the long-term choices of groups.*

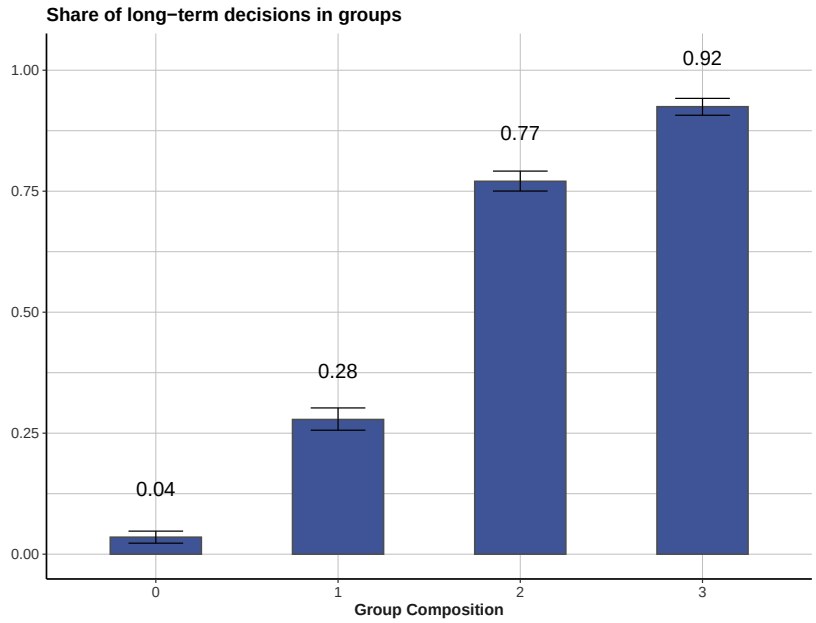


Figure 2.11: Share of long-term decisions in groups depending on group composition pooled for all treatments.

Note: Black whiskers represent bootstrapped 95% confidence intervals. *Group Composition* indicates the number of group members who chose the long-term project in the individual task.

Finally, we take a closer look at the dynamics of the group decisions. We have seen that in mixed group compositions where either one or two subjects are in favor of the long-term project, most often the majority outcome is implemented, meaning that the minority group member switches her decision. Table A2.9 in the Appendix shows the individual characteristics of subjects that are in the minority of a mixed group composition and switch their opinion to either the short-term outcome or long-term outcome, depending on the initial preference. Table 2.6 displays an overview of the results.

Table 2.6: Overview of Switching Behavior of Minorities in Groups

| Group Composition | Individual Choice | Choice in Group | n | Frequent Characteristics of Switcher |
|---|---|---|---|---|
| L-S-S | L | S (switch) | 160 | Impatience, Male, Young, Emotionally |
| L-S-S | L | L (stick) | 67 | Unstable, Not First-Speaker |
| S-L-L | S | S (stick) | 58 | Less Open for Experience, Not |
| S-L-L | S | L (switch) | 206 | First-Speaker |

Note: *L* and *S* indicate a long-term (L) or short-term (S) choice. *Group Composition* indicates the composition of the group by their choice in the individual task. *Individual Choice* and *Choice in Group* show the choices of the minority group member in the individual and group setting, while *n* displays the number of decisions of the corresponding case. *Frequent Characteristics of Switchers* are taken from Table A2.9 in the Appendix and display covariates that increase the probability of switching one's individual choice in the group setting.

In 70.5% of all decisions in which one subject has a preference for the long-term outcome but faces two group members who are in favor of the short-term outcome, the minority group member switches her opinion. Less patient, male, younger and emotionally unstable subjects are more likely to be convinced by the majority to choose the short-term outcome. In 78% of all situations in which the majority of the group is in favor of the long-term project, subjects that initially prefer the short-term project switch their decision. Subjects that are less open to experiences are more likely to switch their opinion.

In addition, we analyzed the communication within each group before each decision in Table A2.9 and Table A2.10 in the Appendix. The chat behavior shows that subjects in the minority of a mixed group composition who write the first message are over 20% less likely to switch their opinion. In general, if subjects write the first message is mainly driven by the degree of extroversion. Comparably, more extrovert subjects write longer messages in general, while women tend to write less.

Lastly, we study if the switching behavior of subjects in a minority of a group is influenced by our treatments when considering the individual characteristics discussed above. When looking at the treatment differences, we see that subjects in the imaginative and intuitive treatment who initially prefer the short-term outcome are less likely to switch their opinion when facing a majority that is in favor of the long-term outcome. This outcome holds when controlling for individual characteristics and the chat behavior (see Table A2.11 in the Appendix).

**Result 7** *Individual characteristics such as patience, gender, and personality traits, as well as the communication behavior influence whether minority group members switch their decisions in groups.*

Altogether, our analysis of long-term decisions in groups demonstrates the importance of individual preferences of group members. This is arguably also often the case in collective decisions taken in politics or the private sector. People first define their individual preferences before discussing a collective decision with their coalition partners or board members. Our results suggest that in order to enhance collective group decisions, it would be beneficial to first guide individuals with mental imagery such that individual future-mindedness is fostered before operating in groups.

### 2.4.6 Robustness Checks

We finally analyze the robustness of the presented results in the individual decision-making setting. We have seen in Model (2) of Table 2.2 and 2.4, that the influence of mental imagery on long-term decision-making and the corresponding mechanism through which mental imagery mainly works, are robust with respect to a variety of covariates. This means that individual and project characteristics themselves do not explain why mental imagery leads to more long-term decisions compared to other mindsets.

#### Experimenter Demand Effect

One could be concerned that subjects in the imaginative treatment might have felt pressured to choose the long-term option, as this might be how they interpreted the implementation intention plan that was given to them in the instructions. To rule out an experimenter demand effect, we asked subjects in the imaginative, intuitive and deliberative treatment about their subjective evaluation of how much they felt influenced and pressured by the respective mindset in their decision-making. The results are shown in Figure 2.12. Figure 2.12A shows the average feeling of being influenced, while Figure 2.12B shows the average feeling of being pressured. Subjects in the imaginative treatment did not feel differently influenced or pressured compared to the other two mindsets. Therefore, we can exclude an experimenter demand effect leading to more long-term decisions in the imaginative treatment.

#### Experiment with General Population

To further check the robustness of our results, we conducted an adjusted version of the experiment online with a representative sample of the German working population based on quotas on gender, age, education and location. In the online experiment, we implemented the same frameworks as in the lab experiment with minor adjustments. First, subjects had to type the corresponding implementation intention plan into an open text field instead of writing it onto a sheet of paper. Next, as mentioned above, subjects in the control treatment were informed about the topic of the study (*"How do people make different donation decisions?"*) which they had to repeat and type in instead of drawing figures. Finally, we added a variation of the imaginative treatment, which we call *Imaginative Vivid*, to test an intensity variation of mental imagery. Compared to the treatment *Imaginative Base* we added one sentence to the imple-

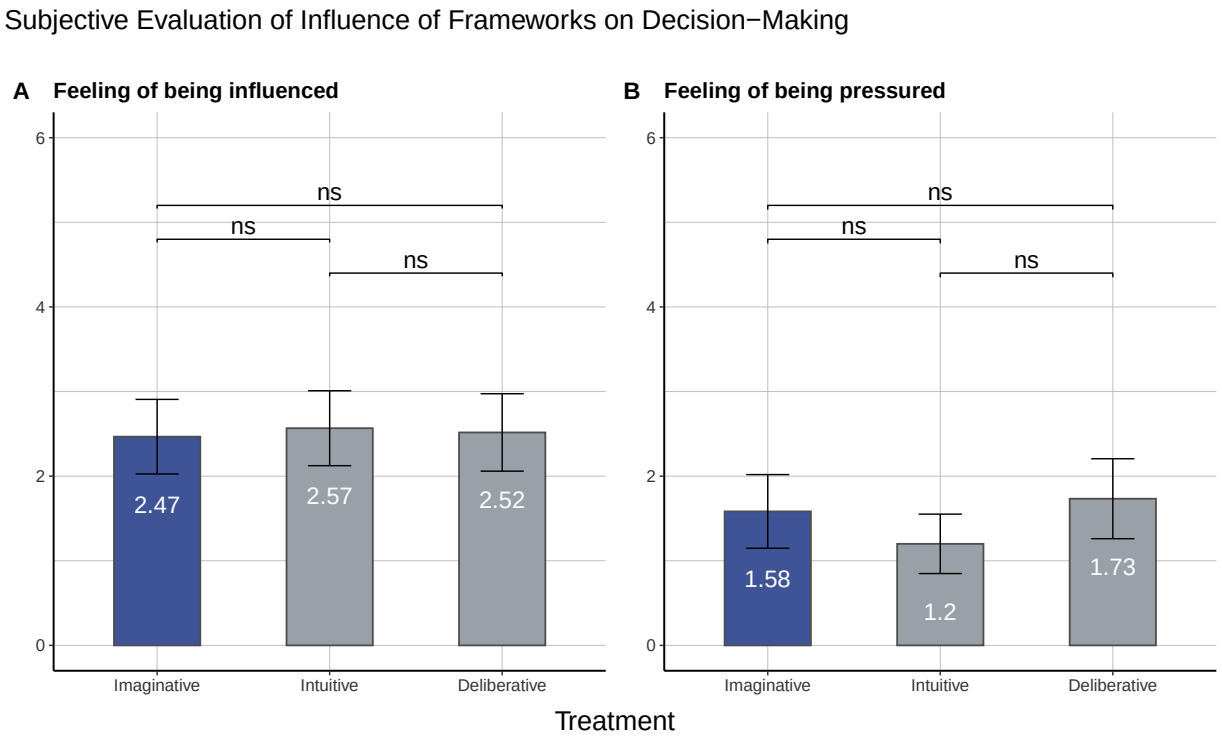


Figure 2.12: Subjective feeling of being influenced (Panel A) or pressured (Panel B) by different mindsets.

Note: Black whiskers represent 95% confidence intervals. Feeling of being influenced / pressured is a self-reported measure and the wording for the elicitation can be found in Table A2.3 in the Appendix. Wilcox Rank-Sum Tests used for treatment comparison.

mentation intention, then saying *"As soon as I start to behave in a routine way, I will tell myself: Imagine what could be! Close your eyes and create a vivid image - what does it look like and how does it feel?"*

Similar to the results in the lab experiment, we see that the response times in the online experiment mimic the induced mindsets (see Figure A2.1 in the Appendix). Subjects in the deliberative treatment took 3.27 min (median = 2.80 min, SD = 1.83 min) on average to complete the investment task, which is significantly more than subjects in the intuitive treatment (mean = 2.5 min, median = 2.33 min, SD = 1.12, $p < 0.001$). In the imaginative treatments subjects took similarly long as in the deliberative treatment (mean = 2.82 min, median = 2.76 min, SD = 1.18 min, $p = 0.230$). This shows that different mindsets can also be successfully induced via a simple reading instruction in an online setting.

Figure 2.13 shows the average share of long-term decisions across the five treatments in the online experiment. Table A2.12 in the Appendix adds the corresponding regression models.

In the online experiment, subjects choose the long-term option less frequently compared to the lab setting, irrespective of the treatment. Still, the treatment effects indicate a similar tendency, although not as strong as in the lab experiment. The deliberative treatment leads to less long-term choices compared to the imaginative treatment ($p < 0.09$). Another purpose of the online experiment was to vary the intensity of mental imagery to see if a more concrete

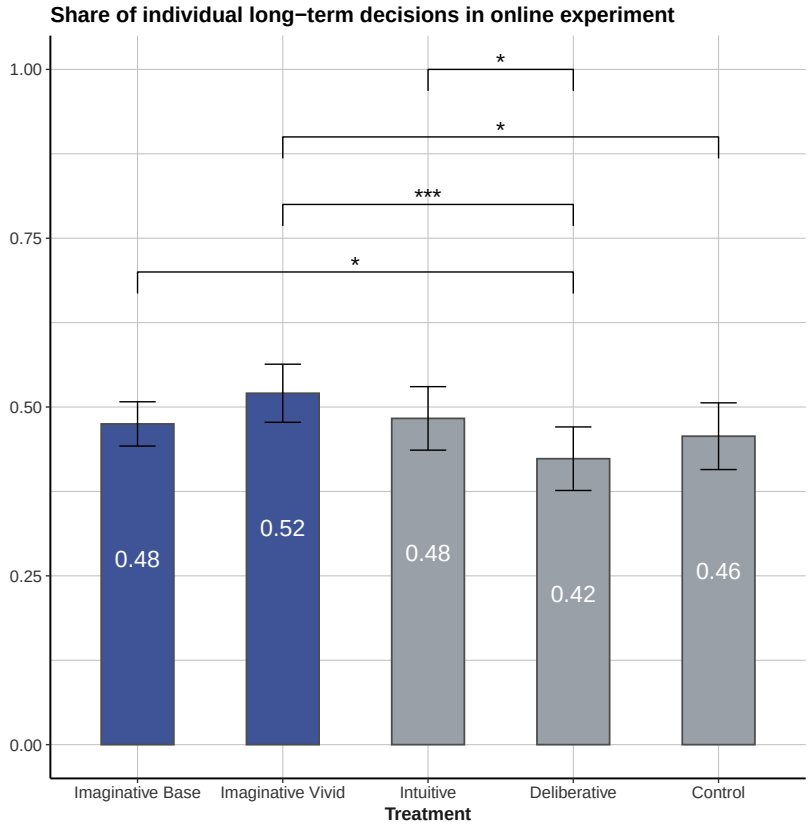


Figure 2.13: Share of long-term decisions in general population sample across treatments.

Note: Black whiskers represent 95% confidence intervals. Wilcox Rank-Sum Tests used for treatment comparison.

visualization can foster future-mindedness even more. We find that subjects in the more intense mental imagery treatment show a share of long-term decisions of 52% which is more than in the deliberative and control treatment (42% and 46%, $p\text{-}values < 0.9$).

Finally, we analyze if the same mechanism of mental imagery applies to a general working population as compared to students. Figure A2.2 in the Appendix shows the adjusted graph of Section 2.4.3 to show how the predicted probability of choosing the long-term option depends on the time horizon in the online experiment. Surprisingly, online subjects' long-term choice does not depend in any way on the time horizon of the long-term project. Neither are long-term projects with a large time horizon chosen less frequently compared to projects with a shorter time horizon, nor is there a treatment effect. This is not what we would expect given our analysis above. In order to check if this result is due to the projects selected or if subjects in the online experiment are not influenced by the time horizon at all, we asked subjects for their last investment decision if they would hypothetically change their decision if the time horizon would change. More specifically, we asked subjects who chose the long-term option, if they would stick to the long-term option if the time horizon would be enlarged by one, two and three years respectively. In contrast, subjects who chose the short-term project were asked if they would switch to the long-term option if the time horizon would be shortened by one, two or

three years. Regression Table 2.7 shows that indeed, the probability of sticking to the long-term choice decreases if the time horizon is enlarged and vice versa, the probability of switching from the short-term to the long-term choice increases if the time horizon of the long-term project is shortened. This shows that subjects in general care about the time horizon, but that the time horizon of the long-term project does not play a role in their decision-making of the investment task.

Table 2.7: Effect of Changing Time Horizon on Long-Term Choice

| | *Dependent variable:* Choosing Long-term Project Hypothetically | |
|---|---|---|
| | Model (1) Stick to Long-term | Model (2) Switch to Long-term |
| Time Horizon Enlarged | $-0.083^{***}$ | |
| | (0.014) | |
| Time Horizon Shortened | | $0.114^{***}$ |
| | | (0.017) |
| Constant | $0.978^{***}$ | $-0.078^{***}$ |
| | (0.028) | (0.029) |
| Observations | 504 | 459 |
| $R^2$ | 0.030 | 0.068 |
| Adjusted $R^2$ | 0.028 | 0.066 |

Note: Linear probability models with binary long-term choice as dependent variable. Model (1) includes decisions in which subjects are asked if they would stick to the long-term option they chose before. Model (2) includes decisions in which subjects are asked if they would switch to the long-term option when they chose the short-term option before. *Time Horizon Enlarged* and *Time Horizon Shortened* can take values from 1-3 and indicate the hypothetical change of the time horizon of the long-term project. Standard errors are clustered at the individual level.
$^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01

Looking at the individual characteristics of subjects, we see little effect on long-term choices, similar as in the lab experiment (see Table A2.13 in the Appendix). Older subjects are less likely to choose the long-term option. This effect was not present in the lab experiment due to the low age span of students, but is in line with the literature showing a negative association between age and future-mindedness (e.g., Busemeyer (2024)).

The findings of the online experiment suggest several insights. First, there is a tendency that mental imagery does also play a role in an online setting with a general population sample, although long-term preferences seem to be weaker in general. Next, the intensity of mental imagery can be varied and might lead to an even bigger shift towards long-term choices. Furthermore, there seems to be an effect of how related subjects are to the corresponding projects. The projects that we selected for the online experiment stem from different organizations across Germany. It is very likely that most of the subjects are not directly related in any way to the organizations. In contrast, in the lab experiment, students had to decide about projects directly

related to their organization (university). Even if many students are not involved in any of these projects, they are related by the institution itself. This might be a potential explanation of why the time horizon of the long-term projects does not influence subjects' long-term choice in the online experiment compared to the lab experiment.

Finally, the set-up of a proper control treatment is important to test the impact of mental imagery. In the lab experiment, we did not hypothesize the control treatment to perform as similar as the imaginative treatment. A potential explanation could be the unintended similarity of the two treatments. When designing the control treatment, we followed the implementation of Barrafrem and Hausfeld (2020) in which subjects had to draw geometric figures that were shown on the screen onto a sheet of paper. Since subjects in the other treatments had to write the respective implementation plans of the mindsets onto a sheet, we wanted to mimic the action in the control treatment without inducing a mindset. However, drawing pictures in the control treatment might have triggered an unintended but similar guidance of thought as the mental imagery framework. Studies in cognitive and social psychology have shown that subjects often either think in verbal descriptions or in images (e.g. Mathews et al. (2013)). Therefore, drawing pictures might have triggered an image-based way of thinking. The results in the online experiment partially confirm this reasoning. The control treatment in the online experiment, mimicking a pure informational treatment, presumably without inducing any image-driven thought, leads to less long-term choices compared to the vivid implementation of mental imagery ($p < 0.09$).

## 2.5 Conclusion

Tackling short-termism – a tendency to prioritize short-term outcomes at the expense of long-term objectives – poses a severe challenge in corporate business, politics and our society. In this paper, we test an intervention of mental imagery to overcome short-termism. Mental imagery is a form of thinking about future outcomes by creating concrete and vivid images in the mind's eye. In two experimental settings, one with university students and one with a representative sample of the German working population, we let people decide between real short-term and long-term investments of their own organization or across nation-wide organizations. Crucially, the university students have minimal chances of benefiting directly of long-term investments, as these have a large time horizon of up to a decade and mainly serve the next generation. The set-up mimics the problem of short-termism, as the decision-makers have to decide in the present about long-term outcomes outlasting their own life cycle.

We impose mental imagery with the help of standard reading instructions to put subjects into an imaginative mindset before making the investment decisions. We compare the effect of mental imagery to an intuitive and deliberative mindset, which incorporate other forms of processing decision-making problems. Our results show that mental imagery leads to more long-term decisions compared to the other mindsets. The effect is driven by the perception of the time horizon. Mental imagery leads to more long-term choices for investments with a

particularly long time horizon. By creating visual images, long-term projects in the far future appear more concrete, the psychological distance imposed by the long time horizon is reduced, and more long-term choices are made. The effect of mental imagery is heterogeneous and related to optimistic views about one's own future life satisfaction. In order to study collective long-term choices, as in board meetings or coalition discussions, we let subjects additionally decide in groups about short-term and long-term investments. Long-term decisions in groups heavily depend on the individual preferences of the group members. In order to generate more long-term decisions in groups, it is important to first individually shift people's preferences from short-term to long-term investments before deciding together in a group.

Our study design allows testing the immediate effect of mental imagery on long-term decisions. However, it remains open how long-lasting effects of mental imagery can be achieved with simple and scalable interventions. Our results suggest that the intensity of mental imagery matters for future-mindedness. Combined with findings from the literature showing that mental imagery is an individual skill that can be trained (e.g., Ashraf et al. (2021)), future research could focus on fostering future-mindedness persistently. Furthermore, while our experiments shed some light on the mechanism of how mental imagery leads to more long-term decisions, it remains to future research to study the mechanisms more thoroughly. We find that the perception of the time horizon of long-term projects in the far future is affected. However, this only holds for subjects that choose among projects within their own organization and not across nation-wide organizations. This leaves room to study the role of one's own relatedness towards future outcomes in mental imagery.

Taken together, we suggest that a simple and short reading intervention of mental imagery (compared to other decision-making mindsets) can mitigate short-termism by fostering future-mindedness. Our intervention is simple and cost-efficient to implement and was validated by two different experimental samples. Shifting individuals' decisions to more long-term outcomes, then in turn leads to more collective long-term investments. As Elise Boulding states it in the quote at the beginning of this paper, thinking about the future requires the capacity to imagine the future. One way of strengthening our imaginative capacity is by using mental imagery.

# 2.6 Appendix

Table A2.1: List of project pairs with corresponding time horizons in lab experiment.

| ID | Project Domain | Short-term | Time | Long-term | Time | Real? |
|---|---|---|---|---|---|---|
| 1 | technical | Bike Garage | 1 yr. | Fabrication Laboratory | 9 yrs. | Yes |
| 2 | cultural | Media Group | 1 yr. | Open Space Exhibition Hall | 10 yrs. | Yes |
| 3 | educational | Student Lab | 1 yr. | Media Studio | 9 yrs. | Yes |
| 4 | social | Café Mondial | 1 yr. | creative.together Lounge | 10 yrs. | Yes |
| 5 | social | Amnesty Group (current project) | 1 yr. | Amnesty Group (future workshop) | 4 yrs. | Yes |
| 6 | social | Daycare Playing Tools | 1 yr. | Daycare Playground | 7 yrs. | No |
| 7 | cultural | Uni Theater (next play) | 1 yr. | Uni Theater (new stage) | 5 yrs. | No |
| 8 | technical | Skate park | 1 yr. | Sailing Storage Building | 8 yrs. | No |
| 9 | educational | Botanical Garden (current project) | 1 yr. | Botanical Garden (tropical house) | 6 yrs. | No |

Table A2.2: List of project pairs with corresponding time horizons in online experiment.

| ID | Project Domain | Organization | Short-term | Time | Long-term | Time |
|---|---|---|---|---|---|---|
| 1 | social | Christopherushilfe e.V. | child care beds | 1 yr | reconstruction main building | 9 yrs. |
| 2 | environ. | Heinz Siemann Stiftung | water supply | 1 yr. | expansion of biotope | 7 yrs. |
| 3 | cultural | Kunst Hilft Geben | workshop / concerts | 1 yr. | integration apartments | 10 yrs. |
| 4 | technical | Deutsche Stiftung Denkmalschutz | Berlin Cathedral | 1 yr. | Kaiser Wilhelm Church | 6 yrs. |
| 5 | sports | Bayerische Sportstiftung | sports material | 1 yr. | job placements | 8 yrs. |
| 6 | educational | Hänsel+Gretel | project "Echt Klasse" | 1 yr. | project "Not-Insel" | 4 yrs. |
| 7 | educational | Deutsche Kinderkrebsstiftung | training / workshop | 1 yr. | research project | 5 yrs. |

Table A2.3: Self-reported measures in lab and online experiment.

| Topic | Measured Variable | Question Text |
|---|---|---|
| Time Preference | Patience | "Are you generally someone who is impatient, or someone who always has a lot of patience?" |
| Risk Preference | Risk-Seekingness | "I am: very risk-avers / very risk-seeking" |
| Social Preference | Altruism | "I am willing to donate to good causes without expecting anything in return." |
| Social Preference | Reciprocity | "If someone does me a favor, I am ready to return it." |
| Social Preference | Trust | "I assume that most people can be trusted." |
| Social Preference | Trust | "I believe that one should not simply trust people and that one cannot be too careful." |
| Future Orientation | Goal Achievement | "When I want to achieve something, I set goals and consider the means by which I can achieve them exactly." |
| Future Orientation | Deadlines | "Meeting tomorrow's deadlines and completing other necessary tasks takes priority over having fun the night before." |
| Future Orientation | Continuous work | "I complete projects on time by steadily working on them." |
| Life Satisfaction | Present | "All in all, how satisfied are you with your life?" |
| Life Satisfaction | Future | "All in all, how satisfied will you be with your life in 10 years?" |
| Framework | Influence | "How much have you felt influenced by the if-then plan?" |
| Framework | Pressure | "How much did you feel pushed by the if-then plan?" |
| Projects | Involvement | "How often do you use the services offered by the listed institutions at the University of Konstanz, or how often are you involved in these institutions?" |
| Projects | Excitement | "How enthusiastic are you about the listed projects or institutions at the University of Konstanz?" |

Note: Translated text into English. All variables are measured on a 7- or 5-point scale.

Table A2.4: Composition of Representative Sample of German Working Population in Online Experiment

| Variable | Outcome | Share |
|---|---|---|
| Gender | Female | 48.5 % |
| Education | No degree (yet) | 0.4 % |
| | Lower level | 12.5 % |
| | High-school | 58.9 % |
| | Bachelor + | 28.2 % |
| Age | < 20 | 0.8 % |
| | 20 - 29 | 16.0 % |
| | 30 - 39 | 21.5 % |
| | 40 - 49 | 19.6 % |
| | 50 - 59 | 27.4 % |
| | 60 - 64 | 10.0 % |
| | 65 + | 4.7 % |
| Nationality | Born in Germany | 94.9 % |
| | Parents born in Germany | 89.6 % |
| Family Status | Single | 38.0 % |
| | Married / Relationship | 50.1 % |
| | Divorced | 8.8 % |
| | Separated | 1.2 % |
| | Widowed | 1.8 % |
| Employment | Full-Time | 63.6 % |
| | Part-Time (> 50 %) | 18.8 % |
| | Part-Time (< 50 %) | 5.1 % |
| | Self-Employed | 7.6 % |
| | Mini-Job | 3.1 % |
| | In Training | 1.8 % |
| Employment (years) | < 1 yr. | 0.8 % |
| | 1 - 3 yrs. | 5.3 % |
| | 3 - 6 yrs. | 8.8 % |
| | 6 - 10 yrs. | 9.6 % |
| | 10 - 15 yrs. | 14.7 % |
| | 15 - 20 yrs. | 12.9 % |
| | > 20 yrs. | 47.6 % |
| Employment (area) | Finance & Insurance | 7.4 % |
| | Construction | 4.1 % |
| | Services | 7.8 % |
| | IT | 7.2 % |
| | Education | 8.2 % |
| | Health | 11.7 % |
| | Consumption | 8.0 % |
| | Real Estate | 2.7 % |
| | Industry | 10.4 % |
| | Agriculture | 1.4 % |
| Employment (leader) | Leader | 35.8 % |

Table A2.5: Robustness of Treatment Effect on Long-Term Choice in Lab Experiment

| | Dependent Variable: Choose Long-term Project |
|---|---|
| Intuitive | −0.057** |
| | (0.028) |
| Deliberative | −0.054* |
| | (0.029) |
| Imaginative | 0.010 |
| | (0.027) |
| Impatience | 0.001 |
| | (0.006) |
| Risk-Seekingness | −0.008** |
| | (0.004) |
| Prosociality | −0.028 |
| | (0.049) |
| Goal Achievement | −0.002 |
| | (0.009) |
| Deadline | 0.003 |
| | (0.008) |
| Continuous Work | 0.003 |
| | (0.007) |
| Future Optimism | 0.023*** |
| | (0.008) |
| Extroversion | −0.003 |
| | (0.007) |
| Conscientiousness | 0.010 |
| | (0.009) |
| Openness To Experience | 0.009 |
| | (0.008) |
| Agreeableness | 0.007 |
| | (0.008) |
| Emotional Stability | −0.007 |
| | (0.008) |
| Female | −0.011 |
| | (0.021) |
| Age | −0.001 |
| | (0.003) |
| Semester | −0.0003 |
| | (0.004) |
| Income | 0.000 |
| | (0.000) |
| Educational Domain | −0.021 |
| | (0.035) |
| Social Domain | −0.071** |
| | (0.029) |
| Technical Domain | −0.182*** |
| | (0.032) |
| Real Long-term Project | 0.040 |
| | (0.028) |
| Existent Long-term Institution | 0.190*** |
| | (0.033) |
| Constant | 0.485*** |
| | (0.109) |
| Observations | 2,097 |
| $R^2$ | 0.050 |
| Adjusted $R^2$ | 0.039 |

Note: Linear probability model with binary long-term choice as dependent variable. The first three covariates are indicators for the respective treatment. *Impatience*, *Risk-Seekingness* and *Prosociality* are incentivized measures of time, risk and social preferences (see Section 2.2.2). *Goal Achievement* until *Future Optimism* are self-reported measures elicited according to Table A2.3. *Extroversion* until *Emotional Stability* indicate personality traits according to a BIG-5 questionnaire. *Educational Domain* until *Existent Long-term Institution* are project-specific dummy variables. Standard errors are clustered at the individual level. *p<0.1; **p<0.05; ***p<0.01

Table A2.6: Influence of individual characteristics on long-term choice

| | *Dependent variable:* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Choose Long-term Project | | | | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Impatience | 0.001 | 0.002 | | | | | | | |
| | (0.006) | (0.006) | | | | | | | |
| Robustness Patience | | −0.0003 | | | | | | | |
| | | (0.006) | | | | | | | |
| Robustness Future Plans | | 0.007 | | | | | | | |
| | | (0.006) | | | | | | | |
| Risk-Seekingness | | | −0.008** | −0.008** | | | | | |
| | | | (0.004) | (0.004) | | | | | |
| Robustness Risk-Seekingness | | | | 0.003 | | | | | |
| | | | | (0.006) | | | | | |
| Prosociality | | | | | −0.031 | −0.062 | | | |
| | | | | | (0.051) | (0.052) | | | |
| Altruism | | | | | | −0.002 | | | |
| | | | | | | (0.007) | | | |
| Reciprocity | | | | | | 0.019 | | | |
| | | | | | | (0.013) | | | |
| Trust | | | | | | 0.014** | | | |
| | | | | | | (0.006) | | | |
| Future Orientedness | | | | | | | 0.013* | | |
| | | | | | | | (0.007) | | |
| Future Optimism | | | | | | | | 0.022*** | |
| | | | | | | | | (0.007) | |
| Extroversion | | | | | | | | | −0.005 |
| | | | | | | | | | (0.008) |
| Conscientiousness | | | | | | | | | 0.010 |
| | | | | | | | | | (0.008) |
| Openness To Experience | | | | | | | | | 0.008 |
| | | | | | | | | | (0.008) |
| Agreeableness | | | | | | | | | 0.009 |
| | | | | | | | | | (0.008) |
| Emotional Stability | | | | | | | | | −0.010 |
| | | | | | | | | | (0.008) |
| Constant | 0.531*** | 0.497*** | 0.567*** | 0.560*** | 0.550*** | 0.392*** | 0.481*** | 0.523*** | 0.471*** |
| | (0.021) | (0.052) | (0.019) | (0.029) | (0.031) | (0.080) | (0.031) | (0.010) | (0.060) |
| Observations | 2,160 | 2,160 | 2,160 | 2,160 | 2,160 | 2,160 | 2,160 | 2,160 | 2,160 |
| $R^2$ | 0.00002 | 0.001 | 0.002 | 0.002 | 0.0002 | 0.004 | 0.001 | 0.003 | 0.002 |
| Adjusted $R^2$ | −0.0004 | −0.001 | 0.001 | 0.001 | −0.0003 | 0.002 | 0.001 | 0.002 | 0.0001 |

Note: Linear probability models with binary long-term choice as dependent variable. *Impatience*, *Risk-Seekingness* and *Prosociality* are incentivized measures of time, risk and social preferences (see Section 2.2.2). The remaining covariates are self-reported measures elicited according to Table A2.3. *Future Orientation* is the average score of the three questions related to future orientation. *Future Optimism* is the difference of future and present life satisfaction. *Extroversion* until *Emotional Stability* indicate personality traits according to a BIG-5 questionnaire. Standard errors are clustered at the individual level. *p<0.1; **p<0.05; ***p<0.01

Table A2.7: Influence of socio-demographic characteristics on long-term choice in lab experiment

| | *Dependent variable:* | | | | | | |
|---|---|---|---|---|---|---|---|
| | Choose Long-term Project | | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Female | 0.010 | | | | | | |
| | (0.020) | | | | | | |
| Age | | −0.002 | | | | | |
| | | (0.003) | | | | | |
| Semester | | | −0.001 | | | | |
| | | | (0.003) | | | | |
| Income | | | | −0.00001 | | | |
| | | | | (0.00004) | | | |
| Plan To Leave Uni | | | | | 0.006 | | |
| | | | | | (0.005) | | |
| Involvement LT Project | | | | | | 0.076*** | |
| | | | | | | (0.016) | |
| Excitement LT Project | | | | | | | 0.113*** |
| | | | | | | | (0.017) |
| Constant | 0.529*** | 0.588*** | 0.541*** | 0.540*** | 0.508*** | 0.508*** | 0.299*** |
| | (0.015) | (0.068) | (0.019) | (0.019) | (0.021) | (0.014) | (0.040) |
| Observations | 2,160 | 2,160 | 2,160 | 2,097 | 1,404 | 1,404 | 648 |
| $R^2$ | 0.0001 | 0.0002 | 0.0001 | 0.00004 | 0.001 | 0.011 | 0.066 |
| Adjusted $R^2$ | −0.0004 | −0.0003 | −0.0004 | −0.0004 | 0.001 | 0.011 | 0.065 |

Note: Linear probability models with binary long-term choice as dependent variable. *Plan to Leave Uni* indicates the number of semesters the subject needs to leave the university. *Involvement LT Project* and *Excitement LT Project* are self-reported measures of the involvement in and excitement about the respective long-term project. These measures are elicited according to Table A2.3. Standard errors are clustered at the individual level. *p<0.1; **p<0.05; ***p<0.01

Table A2.8: Influence of treatments on individual characteristics in lab experiment

| | *Dependent variable:* | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Impatience | Risk-Seekingness | Prosociality | Trust | Future-Orientedness | Future-Optimism | Extroversion | Conscientious | Openness To Experience | Agreeableness | Emotional Stability | Involvement LT Project | Excitement LT Project |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) | (13) |
| Intuitive | 0.167 | 0.683 | 0.010 | 0.400 | −0.261 | −0.083 | 0.108 | −0.050 | 0.200 | −0.108 | −0.125 | −0.133* | −0.176 |
| | (0.314) | (0.504) | (0.042) | (0.297) | (0.226) | (0.214) | (0.269) | (0.246) | (0.216) | (0.218) | (0.249) | (0.078) | (0.190) |
| Deliberative | 0.250 | −0.067 | −0.031 | −0.367 | −0.094 | −0.050 | 0.083 | 0.167 | 0.050 | −0.342 | −0.217 | −0.054 | −0.083 |
| | (0.314) | (0.504) | (0.042) | (0.297) | (0.226) | (0.214) | (0.269) | (0.246) | (0.216) | (0.218) | (0.249) | (0.083) | (0.355) |
| Imaginative | 0.283 | 0.833* | −0.022 | −0.167 | −0.250 | 0.017 | 0.158 | −0.067 | 0.392* | −0.317 | −0.183 | −0.059 | −0.111 |
| | (0.314) | (0.504) | (0.042) | (0.297) | (0.226) | (0.214) | (0.269) | (0.246) | (0.216) | (0.218) | (0.249) | (0.091) | (0.158) |
| Constant | 2.850*** | 3.900*** | 0.533*** | 4.450*** | 4.194*** | 0.550*** | 4.225*** | 5.208*** | 4.950*** | 5.400*** | 5.008*** | 0.358*** | 2.324*** |
| | (0.222) | (0.356) | (0.029) | (0.210) | (0.160) | (0.152) | (0.190) | (0.174) | (0.153) | (0.154) | (0.176) | (0.069) | (0.114) |
| Observations | 240 | 240 | 240 | 240 | 240 | 240 | 240 | 240 | 240 | 240 | 240 | 1,404 | 1,296 |
| $R^2$ | 0.004 | 0.021 | 0.006 | 0.030 | 0.008 | 0.001 | 0.002 | 0.005 | 0.017 | 0.014 | 0.004 | 0.005 | 0.004 |
| Adjusted $R^2$ | −0.009 | 0.008 | −0.007 | 0.017 | −0.005 | −0.012 | −0.011 | −0.008 | 0.004 | 0.002 | −0.009 | 0.002 | 0.001 |

Note: Linear probability models with individual characteristics as dependent variables. *Impatience*, *Risk-Seekingness* and *Prosociality* are incentivized measures of time, risk and social preferences (see Section 2.2.2). The remaining dependent variables are self-reported measures elicited according to Table A2.3. *Future Orientation* is the average score of the three questions related to future orientation. *Future Optimism* is the difference of future and present life satisfaction. *Extroversion* until *Emotional Stability* indicate personality traits according to a BIG-5 questionnaire. *Involvement LT Project* and *Excitement LT Project* indicate the involvement in and excitement about the respective long-term project. The independent variables are indicators for the respective treatment.
Standard errors in Model (12) and (13) are clustered at the individual level. *p<0.1; **p<0.05; ***p<0.01

Table A2.9: Individual characteristics of people who changed opinion as minority in groups

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | Switched To Short-term | | Switched To Long-term | |
| | (1) | (2) | (3) | (4) |
| Impatience | 0.040* | | 0.021 | |
| | (0.021) | | (0.017) | |
| Risk-Seekingness | −0.002 | | −0.001 | |
| | (0.011) | | (0.008) | |
| Prosociality | −0.143 | | 0.087 | |
| | (0.126) | | (0.118) | |
| Altruism | 0.011 | | −0.011 | |
| | (0.023) | | (0.017) | |
| Trust | 0.013 | | −0.007 | |
| | (0.017) | | (0.018) | |
| Reciprocity | −0.002 | | −0.001 | |
| | (0.031) | | (0.035) | |
| Future Optimism | −0.007 | | 0.004 | |
| | (0.016) | | (0.022) | |
| Female | −0.158** | | 0.083 | |
| | (0.064) | | (0.058) | |
| Age | −0.032*** | | −0.006 | |
| | (0.010) | | (0.008) | |
| Extroversion | −0.028 | | 0.007 | |
| | (0.023) | | (0.018) | |
| Conscientiousness | 0.038 | | 0.031 | |
| | (0.025) | | (0.021) | |
| Openness To Experience | 0.025 | | −0.042* | |
| | (0.031) | | (0.025) | |
| Agreeableness | 0.022 | | 0.020 | |
| | (0.029) | | (0.021) | |
| Emotional Stability | −0.046** | | 0.002 | |
| | (0.023) | | (0.021) | |
| Spoke First | | −0.248*** | | −0.225*** |
| | | (0.068) | | (0.061) |
| Message Length | | −0.0001 | | −0.0004 |
| | | (0.001) | | (0.001) |
| Constant | 1.282*** | 0.796*** | 0.779*** | 0.876*** |
| | (0.375) | (0.048) | (0.282) | (0.035) |
| Observations | 227 | 225 | 264 | 264 |
| $R^2$ | 0.096 | 0.068 | 0.045 | 0.068 |
| Adjusted $R^2$ | 0.037 | 0.060 | −0.008 | 0.060 |

Note: Linear probability models with binary variables as dependent variables indicating if subject switched opinion to short-term (Model (1) and (2)) or long-term (Model (3) and (4)) in group setting when being in the minority of a mixed group composition. A mixed group composition is given if not all three group members chose the long-term or short-term projects in the individual setting. *Impatience*, *Risk-Seekingness* and *Prosociality* are incentivized measures of time, risk and social preferences (see Section 2.2.2). The remaining dependent variables are self-reported measures elicited according to Table A2.3. *Future Optimism* is the difference of future and present life satisfaction. *Spoke First* indicates if the subject was the first to write a chat message, while *Message Length* indicates the total length of the written messages per decision. Standard errors are clustered at the group level. *p<0.1; **p<0.05; ***p<0.01

Table A2.10: Influence of individual characteristics on chat behavior

| | *Dependent variable:* | |
|---|---|---|
| | Spoke First | Message Length |
| | (1) | (2) |
| Impatience | 0.004 | −0.692 |
| | (0.008) | (1.028) |
| Risk-Seekingness | 0.002 | −0.101 |
| | (0.005) | (0.769) |
| Prosociality | 0.041 | 9.522 |
| | (0.070) | (8.220) |
| Altruism | 0.002 | 1.846 |
| | (0.009) | (1.392) |
| Trust | 0.002 | 0.275 |
| | (0.009) | (1.214) |
| Reciprocity | −0.007 | 0.650 |
| | (0.014) | (2.596) |
| Future Optimism | −0.008 | 2.210 |
| | (0.012) | (1.676) |
| Female | 0.011 | −9.811** |
| | (0.031) | (4.745) |
| Age | −0.001 | −0.126 |
| | (0.003) | (0.809) |
| Extroversion | 0.016* | 4.365*** |
| | (0.009) | (1.619) |
| Conscientiousness | −0.014 | 0.904 |
| | (0.010) | (1.914) |
| Openness To Experience | −0.007 | −1.022 |
| | (0.012) | (2.201) |
| Agreeableness | −0.001 | −0.352 |
| | (0.012) | (2.117) |
| Emotional Stability | −0.002 | −2.464 |
| | (0.011) | (1.635) |
| Constant | 0.390*** | 39.498 |
| | (0.136) | (33.250) |
| Observations | 2,155 | 2,155 |
| $R^2$ | 0.005 | 0.034 |
| Adjusted $R^2$ | −0.001 | 0.027 |

Note: OLS models with variables indicating who writes the first message (Model (1)) and how long the written messages are (Model (2)) as dependent variables. All group decisions are considered. *Impatience*, *Risk-Seekingness* and *Prosociality* are incentivized measures of time, risk and social preferences (see Section 2.2.2). The remaining dependent variables are self-reported measures elicited according to Table A2.3. *Future Optimism* is the difference of future and present life satisfaction.
Standard errors are clustered at the group level. *p<0.1; **p<0.05; ***p<0.01

Table A2.11: Treatment difference in opinion switch of minorities in groups.

| | *Dependent variable:* | |
|---|---|---|
| | Switch To Short-Term | Switch To Long-Term |
| | (1) | (2) |
| Imaginative | 0.069 | $-0.161^{***}$ |
| | (0.098) | (0.062) |
| Intuitive | 0.102 | $-0.098^{*}$ |
| | (0.100) | (0.057) |
| Deliberative | 0.027 | $-0.109$ |
| | (0.089) | (0.071) |
| Constant | $1.286^{***}$ | $1.086^{***}$ |
| | (0.407) | (0.292) |
| Covariates | Yes | Yes |
| Observations | 225 | 264 |
| $R^2$ | 0.160 | 0.124 |
| Adjusted $R^2$ | 0.083 | 0.056 |

Note: Linear probability models with binary variables as dependent variables indicating if subject switched opinion to short-term (Model (1)) or long-term (Model (2)) in group setting when being in the minority of a mixed group composition. A mixed group composition is given if not all three group members chose the long-term or short-term projects in the individual setting. The independent variables are indicators for the different treatments, respectively. The regressions include all covariates listed in Table A2.9.
Standard errors are clustered at the group level. $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01

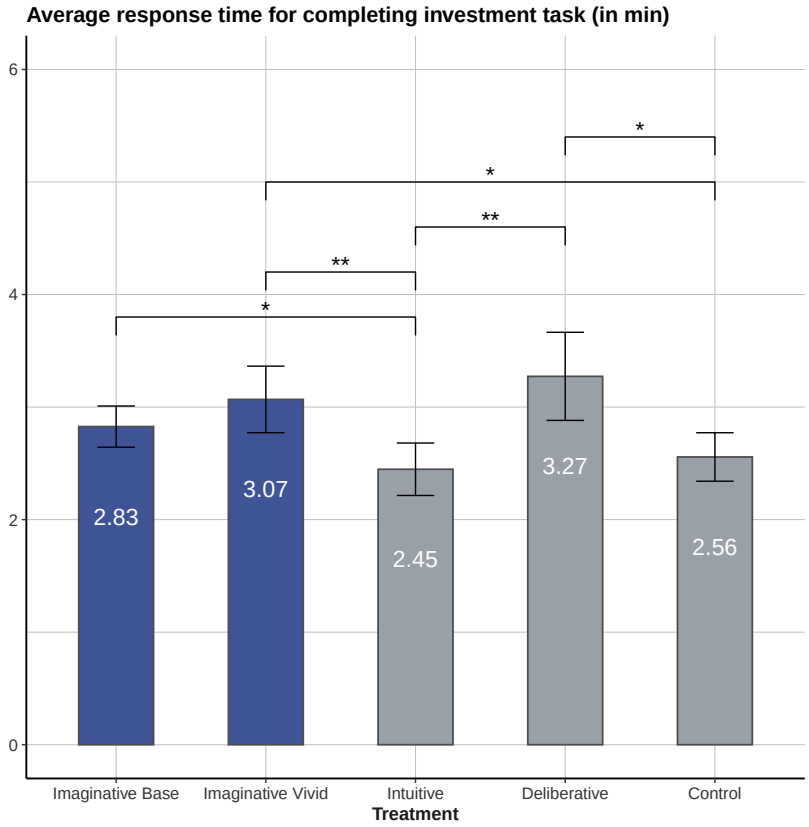


Figure A2.1: Average response times for completing investment task for general population sample in online experiment (in min)

Note: Black whiskers represent 95% confidence intervals. Wilcox Rank-Sum Tests used for treatment comparison.

Table A2.12: Treatment Effect on Long-Term Choice in Linear Probability Models in Online Experiment

| | Dependent Variable: Choose Long-term Project | | | | | |
|---|---|---|---|---|---|---|
| | Model 1 | | | Model 2 | | |
| | Coef. | vs. Imag. Base | vs. Imag. Vivid | Coef. | vs. Imag. Base | vs. Imag. Vivid |
| Imaginative Base | 0.008 | | | 0.008 | | |
| | (0.031) | | | (0.029) | | |
| Imaginative Vivid | 0.046 | p = 0.154 | | 0.059* | p = 0.061 | |
| | (0.034) | | | (0.033) | | |
| Intuitive | 0.018 | p = 0.618 | p = 0.348 | 0.018 | p = 0.691 | p = 0.176 |
| | (0.034) | | | (0.033) | | |
| Deliberative | -0.045 | p = 0.038 | p = 0.003 | -0.048 | p = 0.003 | p < 0.001 |
| | (0.034) | | | (0.033) | | |
| Constant | 0.470*** | | | 0.407*** | | |
| | (0.025) | | | (0.107) | | |
| Covariates | No | | | Yes | | |
| Observations | 3423 | | | 3409 | | |
| $R^2$ | 0.003 | | | 0.095 | | |
| Adjusted $R^2$ | 0.002 | | | 0.084 | | |

Note: Linear probability models with binary long-term choice as dependent variable. The regressors are indicators for the intuitive, deliberative and imaginative treatment (base / vivid) respectively. Covariates include subject- and project-specific characteristics. Second and third column of each model shows the p-values of linear hypothesis tests when comparing the respective coefficient to the coefficient of the imaginative treatment (base and vivid). Standard errors are clustered at the individual level.
*p<0.1; **p<0.05; ***p<0.01

Table A2.13: Effect of Individual Characteristics on Long-Term Choice in Linear Probability Models in Online Experiment

| | *Dependent variable:* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Choose Long-term Project | | | | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Patience | 0.003 | | | | | | | | |
| | (0.007) | | | | | | | | |
| Future Plans | 0.009 | | | | | | | | |
| | (0.007) | | | | | | | | |
| Risk-Seekingness | | −0.006 | | | | | | | |
| | | (0.006) | | | | | | | |
| Altruism | | | 0.003 | | | | | | |
| | | | (0.007) | | | | | | |
| Reciprocity | | | 0.013 | | | | | | |
| | | | (0.011) | | | | | | |
| Trust | | | −0.009 | | | | | | |
| | | | (0.007) | | | | | | |
| Future Optimism | | | | −0.002 | | | | | |
| | | | | (0.008) | | | | | |
| Female | | | | | −0.003 | | | | |
| | | | | | (0.020) | | | | |
| Age | | | | | | −0.002** | | | |
| | | | | | | (0.001) | | | |
| Higher Education | | | | | | | 0.003 | | |
| | | | | | | | (0.021) | | |
| Official Relationship | | | | | | | | −0.030* | |
| | | | | | | | | (0.017) | |
| Leader | | | | | | | | | −0.031 |
| | | | | | | | | | (0.021) |
| Constant | 0.424*** | 0.484*** | 0.423*** | 0.471*** | 0.472*** | 0.547*** | 0.470*** | 0.486*** | 0.482*** |
| | (0.036) | (0.018) | (0.054) | (0.010) | (0.014) | (0.034) | (0.012) | (0.012) | (0.012) |
| Observations | 3,423 | 3,423 | 3,423 | 3,423 | 3,423 | 3,423 | 3,423 | 3,423 | 3,409 |
| $R^2$ | 0.001 | 0.0003 | 0.001 | 0.00002 | 0.00001 | 0.002 | 0.00001 | 0.001 | 0.001 |
| Adjusted $R^2$ | 0.0002 | −0.00001 | 0.0004 | −0.0003 | −0.0003 | 0.002 | −0.0003 | 0.001 | 0.001 |

Note: Linear probability models with binary long-term choice as dependent variable. *Patience* until *Future Optimism* are self-reported measures according to Table A2.3. *Future Optimism* is the difference of future and present life satisfaction. *Higher Education* is a dummy variable indicating an educational degree of Bachelor or above. *Official Relationship* is a dummy variable for marriage or a registered civil partnership. *Leader* is a dummy variable if the subject has leadership responsibilities. Standard errors are clustered at the individual level. *p<0.1; **p<0.05; ***p<0.01

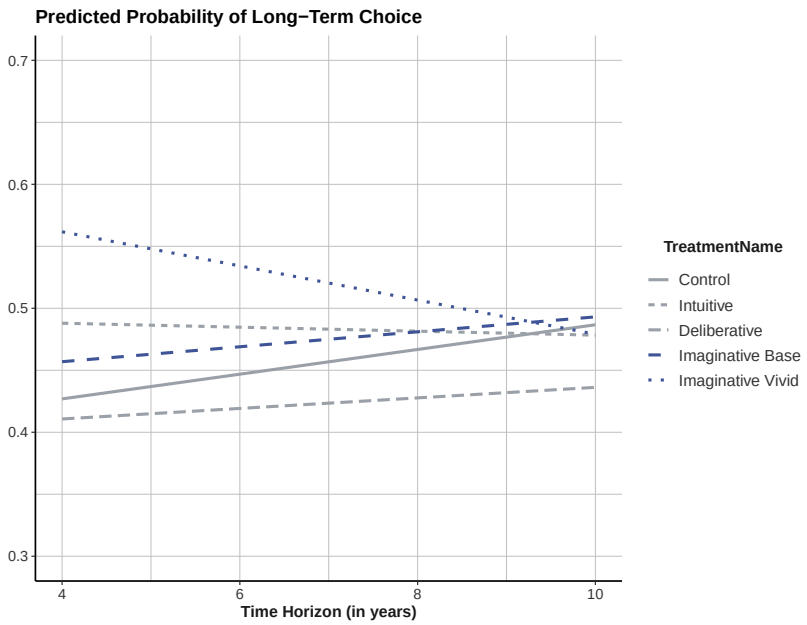


Figure A2.2: Predicted probability of choosing long-term project dependent of time horizon across treatments for the general population in the online experiment.

Note: The graph is based on the linear probability model: $LT\ Choice_i = \beta_0 + \beta_1 * Treatment_j + \beta_2 * Time_i^{LT} + \beta_3 * Treatment_j * Time_i^{LT} + \epsilon_i$, where the dependent variable indicates a binary long-term choice, $Treatment_j$ indicates the respective treatment and $Time_i^{LT}$ denotes the time horizon of the long-term project. Standard errors are clustered at the individual level.

# Chapter 3

# Generative AI Triggers Welfare-Reducing Decisions in Humans

Fabian Dvorak[1,2], Regina Stumpf[3,4], Sebastian Fehrler[5] and Urs Fischbacher[3,4,6]

[1] Centre for the Advanced Study of Collective Behaviour, University of Konstanz (Germany)
[2] Department of Environmental Social Sciences, Eawag (Switzerland)
[3] Department of Economics, University of Konstanz (Germany)
[4] Thurgau Institute of Economics (Switzerland)
[5] SOCIUM, University of Bremen (Germany)
[6] CESifo, Munich, (Germany)

**Abstract**

Generative artificial intelligence (AI) is poised to reshape the way individuals communicate and interact. While this form of AI has the potential to efficiently make numerous human decisions, there is limited understanding of how individuals respond to its use in social interaction. In particular, it remains unclear how individuals engage with algorithms when the interaction entails consequences for other people. Here, we report the results of a large-scale pre-registered online experiment (N = 3,552) indicating diminished fairness, trust, trustworthiness, cooperation, and coordination by human players in economic two-player games, when the decision of the interaction partner is taken over by ChatGPT. On the contrary, we observe no adverse welfare effects when individuals are uncertain about whether they are interacting with a human or generative AI. Therefore, the promotion of AI transparency, often suggested as a solution to mitigate the negative impacts of generative AI on society, shows a detrimental effect on welfare in our study. Concurrently, participants frequently delegate decisions to ChatGPT, particularly when the AI's involvement is undisclosed, and individuals struggle to discern between AI and human decisions.

**Keywords:** Generative AI, Welfare, Efficiency, Economic Games, ChatGPT

**JEL Classification:** D63, D91, O33, I31

**Note:** This is an adapted version of the working paper which can be found here: *arXiv:2401.12773*.

## 3.1 Introduction

Fairness, reciprocity, trust, cooperation, and coordination are fundamental to the proper functioning of human societies. These facets of human behavior support socially desirable outcomes by promoting efficiency through coordination, and stabilizing norms of trust, fairness, and cooperation. Recent advances in the development of generative artificial intelligence have shown the potential to fundamentally change nearly every aspect of life in modern societies. While it is clear that AI has an enormous potential for improving human life (Vinuesa et al., 2020; Lin et al., 2023; Peng et al., 2023; Wang et al., 2023; Yan et al., 2023; Lam et al., 2023; Bi et al., 2023) and increasing economic productivity (Abramoff et al., 2023; Noy and Zhang, 2023), little is known about the consequences of using generative AI in everyday human interaction. In this study, we approach the question of how generative AI will affect human interactions from a welfare perspective. In particular, we ask how the now ubiquitous possibility of using generative AI will affect our everyday social interactions.

Human social behavior depends not only on the expected consequences of our actions for others (Tricomi et al., 2010; Fehr and Fischbacher, 2003), but also on our beliefs about others' behavior (Fehr and Fischbacher, 2004), which jointly explain the variability of culture and social norms (Henrich et al., 2001). If decisions in social interaction are (potentially) delegated to AI, individual's beliefs and preconceptions about the nature of AI, and about the circumstances under which AI systems are employed, will become pertinent factors influencing their decisions (Pataranutaporn et al., 2023).

The focus of this study is *not* how generative AI will behave in social interactions, which has been studied in a recent string of other papers (Bauer et al., 2023; Chen et al., 2023; Guo, 2023; Dargnies et al., 2022). Instead, we investigate *human reactions* to the utilization of generative AI in social interactions, an aspect that has been underexplored thus far.[1] In particular, several fundamental questions regarding human behavior in AI-mediated social interaction remain unaddressed. What are the welfare consequences of human reactions to the use of generative AI in social interaction? Under what circumstances are people willing to delegate decisions to generative AI? What role does the transparency of AI decisions play? And does the potential to personalize generative AI models influence how people react to their use?

We conduct a pre-registered online experiment (N=3,552) to investigate the repercussions of incorporating generative AI into human interaction. In the main experiment, 2,905 participants (mean age: 39.9 years, 49.6% women, 47% college or university degree) engage in direct interactions either with other participants or with the large-language model ChatGPT (OpenAI, 2023) acting on behalf of a human participant with AI support. In each interaction, one of the two participants is supported by ChatGPT. Both participants are affected by the

[1]Research on human-algorithm interaction generally shows that, depending on the design of the algorithm, humans often act more rationally, selfish, and seem less prone to emotional and social responses (Chugunova and Sele, 2022; Köbis et al., 2023), but are nevertheless willing to delegate decisions to algorithms (Candrian and Scherer, 2022). The key feature of our study is that the algorithm acts on behalf of a real participant, who is affected by the consequences of the interaction, which matters for social preferences (von Schenk et al., 2023). More loosely related to our study is the emerging literature investigating humans' willingness to follow algorithmic advice (Kawaguchi, 2021; Logg et al., 2019; Greiner et al., 2024; Zhou et al., 2021).

consequences of the interaction, even if the AI is acting on behalf of the participant with AI support. We provide the same instructions to human participants and the AI and ask for their decisions together with short statements of justification in 5 standard two-player games with direct welfare consequences: the ultimatum game (UG) (Güth et al., 1982), the binary trust game (bTG) (Berg et al., 1995), the prisoner's dilemma game (PD) (Lave, 1962), the stag hunt game (SH), and the coordination game (C). For all five games, a large body of empirical research exists, documenting how humans interact under controlled laboratory conditions (Johnson and Mislin, 2011; Oosterbeek et al., 2004; Mengel, 2017; Dal Bó et al., 2021). To incentivize the decisions, each pair of participants receives the payoffs that result from one randomly selected interaction.

The experiment employs a between-subject design featuring six treatments. In the first treatment condition (*transparent random*), the AI randomly decides on behalf of the participant with AI support with a 50% probability. Participants without AI support make two decisions – one for interacting with a human and another for interacting with AI – rendering the use of AI transparent. In the second treatment condition, the participant with AI support has the option to decide whether to delegate her decision to AI (*transparent delegation*). Meanwhile, the participant without AI support makes two decisions, accounting for each possible outcome of the delegation decision. The third treatment condition is a variation of the delegation treatment, where the participant without AI support cannot condition her decision on the outcome of the delegation decision, rendering the use of AI opaque (*opaque delegation*). For each of these three treatment conditions, we vary whether participants can personalize the decisions AI makes on their behalf at the beginning of the experiment (*personalized*) or not (*non-personalized*, which results in six between-subject treatments overall.

To personalize the decision of AI, participants answered seven binary questions about their own personality before receiving the instructions of the games. The questions followed a simple "A or B?" format with the following pairs of alternatives: Intuition or Thoughtfulness, Introversion or Extraversion, Fairness or Efficiency, Chaos or Boredom, Selfishness or Altruism, Novelty or Reliability, and Truth or Harmony. Participants knew that their answers to the 7 binary preference questions would be used to personalize the AI decisions made on their behalf. They also knew that whenever they interacted with the AI during the experiment, the algorithm would make decisions according to the other participant's preferences, which were elicited through the same questions. We sampled personalized AI decisions for each possible combination of 7 binary preferences ($2^7 = 128$) by prompting ChatGPT to generate decisions made by a person whose preferences were reflected in a particular response pattern (see Section 3.4 for details).

In addition to the main experiment, we test if AI decisions can be detected in social interaction by showing sets of AI decisions and human decisions to 647 human raters (mean age: 39.8 years, 50.1% women, 43.0% college or university degree). Raters receive a bonus payment if they accurately identify the decisions made by the AI. We perform Turing tests by providing human raters with written justifications for the decisions generated by humans and AI.

## 3.2 Experimental Results & Discussion

To quantify the consequences of using AI in social interactions, we use the pre-registered indices shown in Table 3.1 (Dvorak et al., 2023). To construct the indices, we sum up signed normalized decisions in the games, and use the average of the individual averages. In the welfare index, all decisions enter positively.[2] Three additional indices quantify the impact of prosociality, reciprocity, and beliefs about the kindness of the other player, revealing their influence on the underlying welfare effects.

The main hypothesis tests that we present in Table 3.2 were preregistered.[3] The main results are consolidated within the same table, delineating the research questions and their corresponding answers derived from our experimental results for each inquiry. The table specifies the tested variable and the compared treatment conditions.

Table 3.1: Behavioral indices

| Index | Description | Game Decisions (effect) |
|---|---|---|
| Welfare index | Combines all decisions with welfare consequences. | normalized offer in UG (+), normalized minimum acceptance threshold in UG (+), binary trust decision in bTG (+), normalized back-transfer in bTG (+), cooperation in PD (+), cooperation in SH (+), modal choice in C (+) |
| Prosociality index | Combines decisions in which social preferences play a role. Prosocial decisions are defined as decisions that increase the (expected) payoff of the other participant. | normalized offer in UG (+), normalized minimum acceptance threshold in UG(–), normalized binary trust decision in bTG (+), normalized back-transfer in bTG (+), cooperation in PD (+), cooperation in SH (+) |
| Kindness index | Quantifies the beliefs in the kindness of the other player. | normalized binary trust decision in bTG (+), cooperation in PD (+), cooperation in SH (+). |
| Intentions index | Quantifies the role of intentions. | normalized minimum acceptance threshold in UG (+), normalized back-transfer in bTG (+). |

Note: Behavioral indices are created from the normalized game decisions. All indices were pre-registered (https://osf.io/fb7jd).

[2]The only decision for which a positive relation with welfare is debatable is the minimum acceptance threshold in the ultimatum game. Using a positive weight requires that long-term gains of upholding the social norm of utility will outweigh the short-term efficiency losses of rejections. Our results get even stronger if we use a negative weight for the responder decision.

[3]The reported p-values are adjusted for multiple testing using the Holm-Bonferroni method (Holm, 1979). This explains why some of the reported p-values are exactly equal to 1. The alpha level to define statistical significance is 5%.

The key result of our study is that engaging with generative AI triggers human decisions that lead to a decrease in welfare (Question 1). This result emerges consistently across all five experimental games, with significant consequences for participants' payoffs. We further find that participants often delegate their decisions to generative AI, especially, if delegation is opaque (Question 2). Surprisingly, we observe that participants do not alter their behavior when delegation is opaque (Question 3), despite their anticipation of others delegating. Delegation does not crowd-out the social preferences of individuals who are aware that the other person delegated to AI (Question 4). We find that personalizing the AI model does not affect how people respond to it (Questions 5-8). In particular, it does not restore the welfare loss caused by the interaction with AI. We will now delve into each question in more detail. Initially, we present the results from the non-personalized treatments and subsequently highlight the impact of personalization.

Table 3.2: Key findings of the main experiment

| Research question | | Variable | Cond. 1 | Cond. 2 | p-val |
|---|---|---|---|---|---|
| 1. Does interacting with AI decrease welfare? | *Yes* | welfare index | HU: 0.69 | AI: 0.65 | $< 0.001$ |
| 2. Is opaque delegation more frequent? | *Yes* | delegation freq. | TD: 0.38 | OD: 0.42 | 0.006 |
| 3. Does opaque interaction decrease welfare? | *No* | welfare index | HU: 0.69 | UN: 0.71 | 1.000 |
| 4. Does delegation crowd-out prosociality? | *No* | prosociality index | R: 0.55 | D: 0.53 | 0.422 |
| Does personalizing the AI... | | | | | |
| 5. ...restore welfare? | *No* | welfare index | P: 0.66 | NP: 0.65 | 1.000 |
| 6. ...increase delegation? | *No* | delegation freq. | P: 0.39 | NP: 0.40 | 1.000 |
| 7. ...change welfare if delegation is opaque? | *No* | welfare index | P: 0.70 | NP: 0.71 | 1.000 |
| 8. ...change prosocial behavior? | *No* | prosociality index | P: 0.55 | NP: 0.53 | 0.422 |

Note: All hypothesis tests and indices were pre-registered and controlled for multiple testing (https://osf.io/fb7jd/). Results of one-sided t-tests with Holm-Bonferroni corrected p-values (Holm, 1979). Corrected p-values can be equal to 1. HU: decisions directed at human participant; AI: decisions directed at AI; UN: unknown interaction partner (AI or human); TD: transparent delegation; OD; opaque delegation; R: random takeover by AI; D: delegation; P: personalized AI; NP: non-personalized AI.

**Question 1: Does the use of generative AI in human interaction decrease welfare?** When interacting with the AI, participants' decisions lead to a significantly lower level of welfare (M = 0.65, SD = 0.19) than when interacting directly with a human (M = 0.69, SD = 0.17, t(482) = -5.48, $p < 0.001$). This is also true if the other person has intentionally delegated the decision to the AI (AI: M = 0.63, SD = 0.20, HU: M = 0.71, SD = 0.17, t(480) = -10.19, $p < 0.001$) and if the AI is personalized to reflect the preferences of the other person (AI: M = 0.63, SD = 0.20, Human: M = 0.71, SD = 0.17, t(1902) = -15.42, $p < 0.001$). Most of the components of the welfare index contribute to this result.[4] Beliefs in the kindness of the other player are significantly reduced when interacting with AI (M = 0.61, SD = 0.32) compared to interacting with humans directly (M = 0.66, SD = 0.31, t(482) = -3.78, $p < 0.001$). The frequency of the modal choice in the coordination game (earth), which we use as a measure for the predictability of the participants' decision, is significantly lower when interacting with

[4]We pre-registered to analyze the components of the significant main indices.

AI (AI: M = 0.57, SD = 0.50, HU: M = 0.64, SD = 0.48, t(482) = -2.73, p = 0.003). The normalized offer in the ultimatum game, which can be seen as an incentivized measure for equality concerns, is also significantly lower when interacting with AI (M = 0.91, SD = 0.31) compared to direct interaction with a human (M = 0.98, SD = 0.24, t(482) = -5.02, p < 0.001). Reciprocity does not seem to play a role in explaining the welfare losses when interacting with AI (intention index, AI: M = 0.62, SD = 0.19, HU: M = 0.63, SD = 0.19, t(482) = -1.27, p = 0.102).

The reduction in welfare is also reflected in lower levels of expected payoffs of participants when interacting with AI (see bottom panel in Figure 3.1). The simulated expected payoff of all matched decisions that were targeted at the AI (M = 4.74, SD = 0.36) is significantly lower than the one where decisions were targeted at humans (M = 5.06, SD = 0.34, t(1902) = 43.79, p < 0.001). Thus, the involvement of AI in human social interactions leads to an overall loss in expected average payoff of 6%. This discrepancy in expected average payoff varies from a 3% loss in the prisoner's dilemma to a 10% loss in the ultimatum game.

We use linear regression models to explore the extent to which the AI-triggered welfare loss is explained by participants' prior experience with ChatGPT, their socio-economic characteristics, and their attitudes towards AI (models (4)–(6) of Table A3.1). We find that prior experience with ChatGPT does not significantly reduce the welfare loss. Age and positive attitudes towards AI, like the belief that AI is trustworthy and the equality concern in an interaction with AI, mitigate the welfare loss, while the subjective difficulty of predicting AI decisions increases it. Participants that took longer in the experiment show a higher welfare loss.[5]

**Question 2: Is opaque delegation more frequent?** We find that the propensity to delegate the decision to AI increases when delegation is opaque. Comparing the delegation frequency in the treatments with transparent delegation (M = 0.38, SD = 0.29) to the delegation frequency in the treatments with opaque delegation (M = 0.42, SD = 0.29) shows that the propensity to delegate increases by 4 percentage points (t(1944) = 3.14, p = 0.006). This might be explained by the fact that participants anticipate the negative welfare consequences when the use of AI is transparent, and shy away from transparent delegation. Participants' answers to a post-experimental question indicate that transparent delegation is generally considered less appropriate than opaque delegation (transparent delegation (0-4): M = 2.47, SD = 0.97, opaque delegation (0-4): M = 2.61, SD = 0.92, t(1920) = 3.08, p = 0.002).

**Question 3: Does opaque delegation decrease welfare?** Unexpectedly, we find that welfare is not affected if opaque delegation to AI is possible. The decisions of participants who do not know if they interact with a human or AI produce a similar level of welfare (M = 0.71, SD = 0.17) compared to the decisions of participants who know that they interact with a human (M = 0.69, SD = 0.17, t(968) = 1.46, p = 1.000). At the same time, participants

[5]The size of the AI-induced welfare effect is also robust when controlling for socio-economic characteristics of the participants, their prior usage of ChatGPT, and their attitudes towards AI (see models (1)–(3) in Table A3.1).

frequently delegate their decisions to AI (M = 0.42, SD = 0.29) if delegation is opaque and also expect others to use opaque delegation according to a post-experimental question (belief in delegation (0-4): M = 2.85, SD = 0.86).

**Question 4: Does delegation to AI crowd-out prosocial behavior?** Comparing the participants' decisions when interacting with AI in situations in which the decision was delegated to AI with situations in which AI took over randomly allows us to study participants' reactions to delegation. We find that delegation does not crowd out social preferences. Prosocial behavior is unchanged in situations when delegation took place (prosociality index: M = 0.53, SD = 0.19) compared to situations when the AI took over the decision randomly (M = 0.55, SD = 0.20, t(960) = -1.48, p = 0.422). The finding is further substantiated by participants' answers to a question in the post-experimental questionnaire, which indicates that transparent delegation is generally considered appropriate (appropriateness (0-4): M = 2.47, SD = 0.97, comparison to indifference (2), t(945) = 14.98, p < 0.001).

**Questions 5–8: Does personalization of the AI matter?** We find that personalization of AI does not restore the welfare losses that result from the usage of AI in social interaction. The welfare consequences of participants' decisions are similar in the non-personalized random treatment (M = 0.65, SD = 0.19) and the personalized random treatment (M = 0.66, SD = 0.19, t(952) = 0.57, p = 1.000). This also applies in the treatments with opaque delegation (non-personalized: M = 0.71, SD = 0.17, personalized: M = 0.70, SD = 0.18, t(997) = -1.15, p = 1.000). Neither does personalization increase the propensity to delegate to AI (non-personalized: M = 0.40, SD = 0.29, personalized: M = 0.39, SD = 0.29, t(1946) = -0.85, p = 1.000). Responses to our post-experimental question reveal that the utilization of personalization has a noteworthy impact on enhancing the alignment between the AI's decisions and those of the other individual, as indicated by participants' responses (non-personalized (0-4): M = 1.54, SD = 1.10, personalized (0-4): M = 1.94, SD = 1.12, t(2902) = -9.74, p < 0.001). However, participants also indicate that the other person is not adequately represented by the personalized AI (AI adequately represents human (0-4): M = 1.94, SD = 1.12, comparison to indifference (2), t(1451) = -2.07, p = 0.981).

### 3.2.1 Experimental Games

The graphs in the top row of Figure 3.1 show the average decision in each game for three cases: the case where the interaction partner is human (green bars), AI (gray bars) or unknown (blue bars). The bottom graph shows the relative expected payoff differences for two cases where the interaction partner is AI or unknown compared to the case where the interaction partner is human.[6] Whiskers show 95% confidence intervals resulting from non-parametric bootstrapping.

[6]We compare the average payoff for human interactions (where both humans knowingly play against a human) to the average payoff for human interactions in which both humans believe to interact with AI or an unknown interaction partner respectively. We match each participant's decisions targeted at humans, AI or unknown to the respective decisions of all other participants and compute for each participant the mean payoff

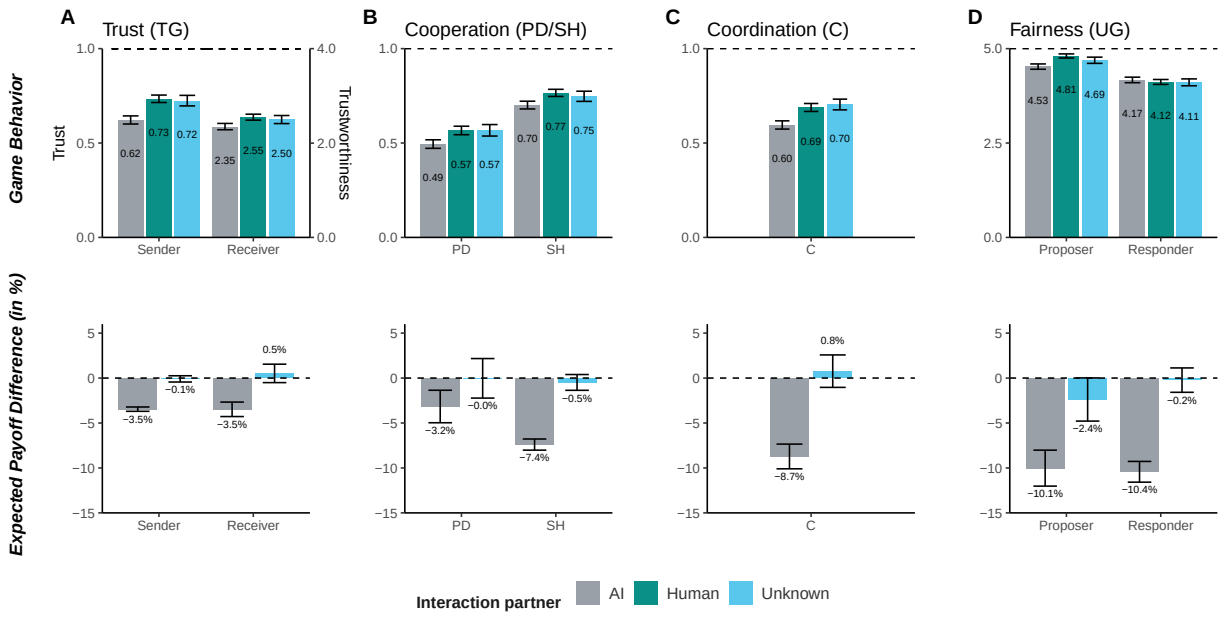


Figure 3.1: Results across four types of games: Trust Game (Panel A), Cooperation Games (Panel B), Coordination Game (Panel C), Fairness Game (Panel D). The color of the bars represent decisions in which the interaction partner is AI (gray bars), human (green bars) or unknown (blue bars).

Top row: average decision in each game. Bottom row: simulated expected payoff difference. Each participant's decisions targeted at humans, AI or unknown is matched to the respective decisions of all other participants and the mean payoff is computed for each participant and each situation resulting from all matched interactions. Expected relative payoff difference to human interactions are reported. Whiskers show bootstrapped 95% confidence intervals.

Comparing the decisions directed to the AI with the decisions directed to other humans, we see welfare-decreasing decisions in all five experimental games. Panel A shows that participants place less trust in the AI (M = 0.62, SD = 0.48) than in humans (M = 0.73, SD = 0.44), and are less trustworthy if the trust decision was made by AI (M = 2.35, SD = 1.50) than when it was made by a human (M = 2.55, SD = 1.38). Both effects together imply a payoff loss of 3.5 percent in each role of the trust game.

Panel B shows that participants cooperate less frequently in the prisoner's dilemma when the other player's decision is taken by the AI (M = 0.49, SD = 0.50) than when taken by a human (M = 0.57, SD = 0.50), which decreases the average payoff in this game by 3.2 percent. The same result applies to the stag-hunt game (AI: M = 0.70, SD = 0.46; HU: M = 0.77, SD = 0.42), leading to a payoff loss of 7.4 percent in this game. Panel C shows that the predictability of participants' actions in the coordination game suffers when interacting with the AI (M = 0.60, SD = 0.49) compared to the human interaction (M = 0.69, SD = 0.46), resulting in a payoff loss of 8.7 percent.

Panel D shows that participants offer less as proposers in the ultimatum game if they know that the decision of the responder is taken over by the AI (M = 4.53, SD = 1.56) compared to

for each situation resulting from all matched interactions. We report the expected mean difference in payoffs, comparing it to the baseline payoff resulting when the interaction partner is human.

when the responder herself takes the decision (M = 4.81, SD = 1.22). When the AI makes the offer in the ultimatum game, participants increase their minimum acceptance threshold and tolerate less negative inequality (M = 4.17, SD = 1.62) compared to when the offer is made by a human (M = 4.12, SD = 1.47). This is the only situation in which the reaction to the AI is potentially welfare increasing in the long run, as participants are willing to uphold the fairness norm at personal costs.[7] At the same time, the larger minimum acceptance threshold has negative consequences for efficiency, as it increases the likelihood of rejections. Combined with the lower offers, the interaction with AI reduces the expected payoffs of the proposer and the responder by 10.1 percent and 10.4 percent, respectively.

The welfare-reducing choices made when interacting with AI are substantiated by participants' answers in the post-experimental questionnaire. Participants think AI is less trustworthy and less cooperative than humans (trustworthiness and cooperation (0-4): AI: M = 2.08, SD = 0.81; Human: M = 2.25, SD = 0.71; testing difference in trustworthiness and cooperation (0): M = -0.17, SD = 0.99, t(2904) = -9.44, $p < 0.001$), which explains reduced trust as sender in the trust game and less cooperation in the prisoner's dilemma and the stag-hunt game when the decision of the other participant is made by AI. Participants also indicate that they care less about equality when AI makes the decision on behalf of the other person (equality concern (0-4): AI: M = 2.98, SD = 1.05; Human: M = 3.26, SD = 0.95, testing difference in equality concern (0): M = -0.28, SD = 0.87, t(2904) = -17.06, < 0.001), which explains lower offers as proposers in the ultimatum game and lower back-transfers as receivers in the binary trust game.

The blue bars in Figure 3.1 show that participants' decisions in situations in which they do not know if they interact with a human or the AI are similar to their decisions directed at humans (green bars). This is the case for every decision in each game, and suggests that, if in doubt, participants behave as if they would directly interact with a human. Consequently, there is no expected payoff difference between human interactions and interactions in which the interaction partner (AI or human) is not known.

### 3.2.2 Delegation Behavior

We now focus on the delegation behavior to AI, which is illustrated in Figure 3.2. The left panel of Figure 3.2 shows that at least 20% of all decisions are delegated to AI in each game-specific situation. This is remarkable because the decisions have consequences. If participants delegate, they do not have to write a justification for the decisions, which creates an incentive to delegate. Additionally, delegation can be used strategically to avoid being blamed for selfish decisions. Decisions in the stag-hunt game are least often delegated (transparent: 24%, opaque: 29%) while the binary trust decision is delegated over 50% of the time (transparent: 51%, opaque:

[7] Our welfare index relies on the notion that the benefits of upholding a social norm of equality will outweigh the efficiency loss in the long run. A higher minimum acceptance threshold is therefore welfare increasing according to our welfare index. It should be noted that modifying our welfare index by considering higher minimum acceptance thresholds as welfare decreasing will make the welfare losses induced by generative AI even stronger. Welfare is also significantly reduced when excluding the ultimatum game from the welfare index.

55%). This may be explained by the fact that cooperation seems to be very attractive in the stag-hunt game (M = 0.81, SD = 0.39) and participants do not want to risk that the AI does not cooperate. Contrasting the frequencies shown in the left panel of Figure 3.2 to participants' answers of the post-experimental questionnaire, we find that participants anticipate the actual delegation behavior well by stating that others' will delegate to AI (belief in delegation (0-4): M = 2.86, SD = 0.84, comparison to indifference (2), $t(1947) = 45.51$, $p < 0.001$).

Figure 3.2 also shows that participants consistently delegate more if delegation is opaque (transparent delegation: M = 0.38, SD = 0.29, opaque delegation: M= 0.42, SD = 0.29, $t(1944) = 3.14$, $p = 0.006$). In all games except the coordination game, opaque delegation is more frequent than transparent delegation (p-values in all situations except C < 0.070), which could indicate that participants anticipate the negative welfare consequences of transparent delegation.

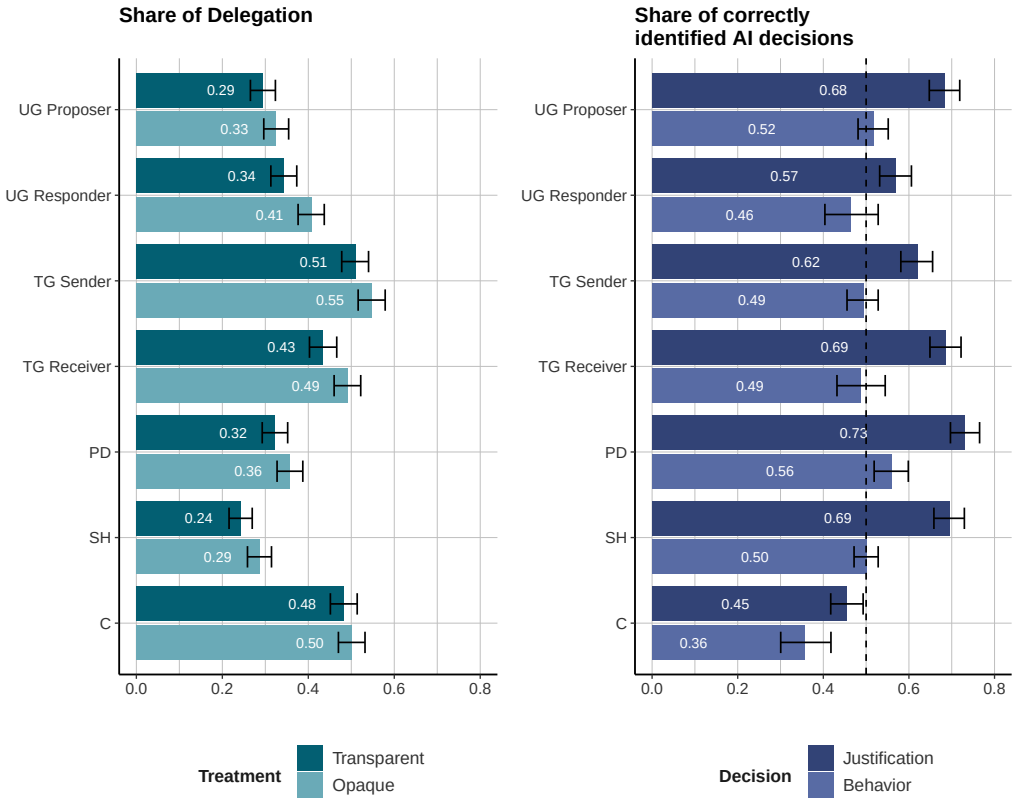


Figure 3.2: Delegation to AI and detectability of AI decisions.

Left: share of delegated decisions in treatments *transparent delegation* (light green bars) and *opaque delegation* (dark green bars). Right: Share of correctly identified AI decisions of human raters when seeing pairs of AI and human decisions with (dark blue bars) and without statements of justification (light blue bars). Whiskers show bootstrapped 95% confidence intervals.

### 3.2.3 Detectability & Turing Tests

We now turn to the question whether people are able to recognize AI decisions or AI-generated justifications for decisions. In each of the seven decisions, humans chose all available options with positive probability, making it fundamentally difficult to distinguish between human and

AI decisions. The right panel of Figure 3.2 shows that the 647 independent human raters who did not take part in the main experiment are generally unable to detect AI decisions. In all games except the prisoner's dilemma, the success rate of human raters is not statistically better than random guessing ($p > 0.05$ for all situations except PD) when presented with two decisions, one of which was generated by the AI (light blue bars). In the prisoner's dilemma, the AI always cooperates, which is more risky, while humans sometimes choose to defect. This makes identification of the AI possible (success rate PD: $M = 0.56$, $SD = 0.30$, $t(212) = 2.86$, $p = 0.002$). In contrast, the success rate of human raters is lower than 50% in the coordination game (two-sided t-test: $p < 0.001$), which shows that humans can be misled by seemingly human-like decisions made by AI.

The dark blue bars in the right panel of Figure 3.2 reveal that the AI does generally not pass the Turing test if raters are additionally shown the written statements of justification for its decision. In all games except the coordination game, the success rate of human raters is significantly higher than chance (p-values for all situations except C $< 0.001$). The AI generated justifications are generally more technical, often referring to the expected payoffs of the alternatives, than human justifications, which justify decisions with simple language. This might stem from the fact that the AI's training data related to the games mainly stems from texts that target a scientific audience and consequentially seems to use the same level of technicality in its justifications. The only exception is the coordination game, for which human raters cannot detect AI generated content even when having access to written justifications. For the coordination game, the AI justifications are less technical and sometimes even a little arbitrary.

## 3.3 Conclusion

Recent advancements in generative AI have paved the way for the convenient delegation of a growing number of decisions to algorithms during everyday social interactions.[8] The findings of our large-scale experiment suggest that this development may have detrimental effects on key underpinnings of human welfare, including fairness, trust, trustworthiness, cooperation, and coordination.

The transparent utilization of generative AI in social interactions triggers welfare reducing decisions in humans. This pattern persists across all five canonical games we investigate, resulting in negative consequences for participants' payoffs.[9] Despite this negative effects, participants consistently opt to delegate their decisions to AI when given the opportunity.

[8]By September 2023, ChatGPT had been used 1.5 billion times, and the use of AI on social media platforms is constantly growing (https://www.similarweb.com/website/chat.openai.com/; last retrieved: 08.12.2023).

[9]The implications of our findings are strengthened by the observation that prior experience with ChatGPT does not mitigate the negative welfare effects. Nearly half of our participants indicate having used ChatGPT at least once before the study. This high proportion of users among our participants, in contrast to the general population, rules out the possibility that our results are influenced by a lack of familiarity with ChatGPT in our sample compared to the broader population. Official reports indicate that 8% of US citizens are ChatGPT users, while this increases to 10% for the population of the UK (https://nerdynav.com/chatgpt-statistics/, last retrieved: 08.12.2023).

This is particularly noteworthy given the widespread availability and often cost-free nature of sophisticated generative AI models. Notably, participants exhibit a higher tendency to delegate their decisions when the use of AI is opaque. This suggests an expectation of potential negative welfare consequences linked to transparent delegation.

Additionally, we discover that opaque delegation to AI can escape detection under specific conditions. In four out of the five games we examine, decisions made by modern generative AI cannot be distinguished when only decisions are reported. In one game, the accuracy of human raters, tasked with identifying decisions secretly delegated to AI, is even lower than chance. These findings underscore the high quality that modern generative AI has achieved in standardized social interactions, such as those explored in our study. However, when raters are provided with justifications for each decision, written by either a human participant or the AI, they can identify AI decisions more frequently than by chance in four out of five games. This highlights that, despite the advancements, some limitations persist in clandestinely employing generative AI in human interaction.

Taken together, our results indicate that recently proposed regulations to safeguard against the detrimental effects of generative AI on society may be inadequate if they do not consider human reactions to such technology. The study suggests that the potential risks arising from AI decisions constitute only part of the problem. Undesirable changes in human behavior could exacerbate these risks. While the EU AI Act[10], one of the first comprehensive AI laws globally, requires transparency regarding content generated by generative AI, our findings suggest that such transparency might have unintended consequences. Despite participants in our experiment believing that others may covertly delegate decisions to AI, we observed no increase in the number of welfare-reducing decisions when the use of generative AI is opaque. This stands in contrast to our findings under transparency. These results suggest a leniency pattern when participants are unaware of whether the other person has delegated decisions to AI. Consequently, the proposition to enhance transparency in AI usage, often advocated to minimize negative effects of generative AI on society, paradoxically results in the most pronounced negative effects on welfare in our study (also see Rahwan et al., 2019; Leib et al., 2024).

An effective regulatory approach, complementing transparency in AI utilization, might involve strengthening public trust in algorithms through independent rigorous testing. To enhance the role of generative AI in social interactions, prioritizing the opening of the black box—rather than merely pointing it out—should be imperative.

[10]https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence; last retrieved: 08.12.2023

## 3.4 Methods & Supplementary Information

**Study type** The study combines online data collection of AI decisions with an online experiment involving human participants. AI decisions and justification statements are elicited using the text interface of ChatGPT (OpenAI, 2023) (Version Jan 30, free research preview) prior to the online experiment. Participants of the online experiment are recruited via Prolific (www.prolific.com) and randomly assigned to one of the six between-subject treatments. Raters are recruited online via Prolific.

**AI decisions** We sampled AI decisions together with a statement of justification for all decisions of the experiment using the text interface of ChatGPT, Jan 30 Version (courtesy of OpenAI). The algorithm received the same instructions the participants receive during the experiment and therefore knows the consequences of the actions for both players (in experimental currency units (ECU)).

We sampled personalized AI decisions for each possible combination of 7 binary preferences ($2^7 = 128$). The algorithm is instructed to act like a person who has preferences defined by a specific pattern of answers to the 7 binary questions. To communicate the preferences of this person to the algorithm, we use the following wording: "Pretend you are a person who prefers intuition over thoughtfulness, introversion over extraversion, fairness over efficiency, chaos over boredom, selfishness over altruism, novelty over reliability, and truth over harmony. Only answer the question at the end. Do NOT explain your answer." For non-personalized decisions, we use the sentence: "Pretend you are a person. Only answer the question at the end. Do NOT explain your answer." After the introductory sentence, the algorithm receives information about the game and the decision.

After the algorithm made the decision, we asked for a statement of justification with the sentence: "Please provide short, informal justification for the decision." Before moving to the next decision, we deleted all prior conversations and restarted the text interface. In rare cases, in which the algorithm did not provide a valid answer or justification, we deleted the prior conversation, restarted the text interface, and repeated the procedure outlined above.

**Experimental games** Participants of the online experiment play five incentivized economic games. The payoffs are in Experimental Currency Units (ECU) and are converted to British pounds for payment.

Ultimatum game (UG)
UG proposer: Endowment of 10 ECU. Offer (0-10 ECU)
UG responder: Minimum acceptable offer (0-10 ECU)

Binary trust game (bTG)
bTG sender: Endowment of 5 ECU. Binary decision to send 2 ECU (tripled).
bTG trustee: Endowment of 5 ECU. Return in case of trust (0-6 ECU)

Prisoner's dilemma game (PD)
Both C, both get 5 ECU. Both D, both get 3 ECU.
You C and the other D, you get 1 ECU and the other gets 8 ECU.
You D and the other C, you get 8 ECU and the other gets 1 ECU.

Stag-hunt game (SH)
Both C, both get 8 ECU. Both D, both get 4 ECU.
You C and the other D, you get 1 ECU and the other gets 5 ECU.
You D and the other C, you get 5 ECU and the other gets 1 ECU.

Coordination game (C)
The five options are:
mercury, venus, earth, mars, saturn
Same option, 5 ECU. Different options, 2 ECU.

For the second-mover decisions in the ultimatum game and in the binary trust game, we use the strategy method and ask participants for the minimum acceptable offer and the conditional back-transfer in the case of trust. In total, there are 7 different situations in the online experiment (UG sender, UG responder, bTG sender, bTG trustee, PD player, SH player, C player).

Participants encounter each situation twice, first as the person with AI support and then as the person without AI support. In case of AI support, depending on the treatment, the participants either have to select a decision for the case that AI does not take over, or they have to decide whether to delegate the decision to AI. Without AI support, depending on the treatment, the participants either have to select a decision for each potential and known interaction partner - AI or human - or select one decision for an unknown interaction partner.

**Experimental treatments** The experiment uses a between-subject design with six treatments. In the first treatment condition, AI makes the decision of the participant with AI support with 50% probability (*transparent random*). Participants without AI support make two decisions, one for the case of interacting with a human, and one for the case of interacting with AI – which makes the use of AI transparent. In the second treatment condition, the participant with AI support can choose if she wants to delegate her decision to AI (*transparent delegation*). In this condition, the use of AI is also transparent as the participant without AI support makes two decisions, one for each outcome of the delegation decision. The third treatment condition is a variant of the delegation treatment in which participants without AI support cannot condition their decision on the outcome of the delegation decision, which makes the use of AI opaque (*opaque delegation*). For each of these three treatment conditions, we vary whether participants can personalize the decisions AI makes on their behalf at the beginning of the experiment (*personalized*) or not (*non-personalized*). This results in six between-subject treatments overall to which participants are randomly assigned:

1. TRN: transparent random non-personalized
2. TRP: transparent random personalized
3. TDN: transparent delegation non-personalized
4. TDP: transparent delegation personalized
5. ODN: opaque delegation non-personalized
6. ODP: opaque delegation personalized

**Monetary incentives** After a participant has completed all 14 decisions, she is randomly matched to another participant from the same treatment. For each pair of participants, the experimenter randomly selects one of the 14 interactions, which determines the game and the players' roles in the game, and randomly determines which participant is supported by AI. If the treatment involves the possibility of random takeover by AI (TRN, TRP), the experimenter randomly determines if AI makes the decision on behalf of the participant with AI support. The payoffs of both participants are then determined using the decisions of both players (human or AI) in the selected interaction. There is no feedback between interactions.

**Personalization of AI** In the treatments with personalization, we use participants' answers to 7 binary preference questions to personalize AI decisions made on their behalf, which results in 128 possible answer patterns. Each unique pattern of answers reflects a certain personality of the AI. The questions follow the simple format "A or B?" and are: intuition or thoughtfulness, introversion or extraversion, fairness or efficiency, chaos or boredom, selfishness or altruism, novelty or reliability and, truth or harmony.

Participants know the purpose of the questions - that their answers will influence the decisions AI makes on their behalf in the online experiment. Participants do not know the instructions of the games at the time when answering the binary questions, which limits (but does not exclude) strategic personalization. At the time when participants personalize the AI, they only know that they will interact with other participants and that their decisions will have consequences for both participants.

**Experimental procedures** All participants receive general instructions with information about the study and the experimental procedures, and are subsequently asked to give informed consent. Participants then receive detailed information about the AI model used in the online experiment. We provide examples of model output in tasks unrelated to the economic games studied in the experiment. We then ask for participants' prior experience with ChatGPT. In the treatments with personalized AI, we provide an additional example for the effect of personalizing the response of ChatGPT. Participants are then asked 7 binary questions in the simple format "A or B?" to personalize decisions made by AI on their behalf. This happens before participants receive the instruction for the games to limit the possibility of strategic personalization.

Participants receive general instructions on how the interaction with the AI model works (treatment specific but not game specific) and answer several control questions, including two attention checks. Participants then receive the instructions for the first game. We use standardized instructions for all games (Thielmann et al., 2021; Mehta et al., 1994) with minor modifications for SH and C. Participants are informed about their role in the game and are asked for their decision(s) with AI support.

For all decisions except those that are delegated to AI, participants write a short statement of justification immediately after making the decision. Participants know that the justifications will not be shown to other participants during the experiment. Participants are subsequently asked for their decision(s) as participant without AI support and their statement(s) of justification. If the game has more than one player role, the procedure is repeated for the other role. No feedback about the decisions of other participants or the AI decisions is provided between the interactions.

Participants with AI support make one decision in the treatments with random AI takeover (TRN, TRP) for the case that AI does not make the decision. In all treatments with delegation (TDN, TDP, ODN, ODP) participants with AI support first decide whether they would like to delegate and are subsequently only asked for their decision if they do not delegate to AI. Participants in the transparent treatments (TRN, TRP, TDN, TDP) make two decisions in the interactions without AI support, one for the case of interacting with a human and, one for the case of interacting with AI. Participants in the opaque delegation treatments (ODN, ODP) make one unconditional decision in interactions without AI support.

After a participant has completed all games in a randomized order, we ask participants for their age, gender, and education level. We also ask each participant to rate the predictability and kindness of AI, the importance of equality and intentions when interacting with AI, the resemblance of AI decision to human preferences, and in the delegation treatments the belief in and the appropriateness of delegation to AI on a 5-point Likert scale. After the experiment, each participant is matched to another participant from the same treatment by the experimenter. The experimenter randomly selects one of the 14 interactions for each pair, and the matched participants receive the payoffs resulting from this interaction in addition to a flat payment for participation.

**Detectability & Turing Tests** For each treatment condition, we create several collections of decisions, each consisting of 7 human decisions and 7 AI decisions. To generate the collections, we filter out all decisions with statements revealing that (1) a human made the decision or (2) AI made the decision. This includes, for example, all human decisions with statements that reference information not available to AI (e.g., information about the general procedure of the experiment, the fact that the participant is interacting with AI) and AI statements revealing AI generated decisions. We also filter out human statements with less than 3 words and correct the selected human statements for typos or grammatical errors.

For the data of four treatments TRN, TRP, TDN, and TDP, each rater compares the 7 decisions a participant made without AI support targeted at a human to the 7 AI decisions

generated for the same participant in the same interaction and indicates for each situation which decision was generated by AI. All decisions are presented together with the corresponding statements of justification written by the participant or written by AI. The rater receives a bonus payment if her rating in one randomly selected situation is correct.

For the data of the treatments ODN and ODP, the raters compare decisions of participants with AI support who do not delegate their decision to the decision generated by AI for these participants in the same situation. We collect 14 ratings from each rater. Each rater reviews 7 decisions, one from each of the 7 situations in the online experiment. First, we present a pair of decisions without the written statement of justification and ask the rater to decide which decision AI made. Then, we reveal the written statements of justification and give the rater the opportunity to revise her rating. At the end, one situation and one of the two ratings is randomly selected, and the rater receives a bonus payment if the rating is correct.

For the Turing tests (exploratory hypotheses H9 and H10), we are interested in the share of correctly identified AI decisions. In the treatments ODN and ODP we post-hoc randomize the positioning of the AI decisions in situations in which the AI and the human actions were identical to exclude any ordering bias. We do so by setting the outcome variable of correctly identifying the AI decision to 0.5. We also analyze participants' Likert-scale ratings about the nature of human-AI interaction and socio-demographic variables collected after the experiment. We also measure response times and how often a participant leaves or switches tabs in the browser (or cannot see the browser window) during the online experiment.

**Ethical approval and informed consent** The online experiment was carried out in accordance with the regulations of the ethics committee of the University of Konstanz. All experimental protocols were approved by the institutional review board of Gesellschaft für Experimentelle Wirtschaftsforschung (GfEW). Institutional Review Board Certificate No. Ja89kRho (https://gfew.de/ethik/Ja89kRho). Informed consent was obtained from all participants prior to participation.

**Sample size** The target sample size was 3000 experimental subjects in the main experiment (500 per treatment) and 600 human raters (100 per treatment) for the Turing tests. We reached a sample size of 2905 experimental subjects (mean age: 39.9 years, 49.6% women, 47.1% college or university degree) and 647 human raters (mean age: 39.8 years, 50.1% women, 43% college or university degree). We estimated the required sample size of each treatment based on a simulation. For each of our main hypotheses (H1-H8), we simulated data with effects of 5 ppt in the direction of the alternative hypothesis for each decision affected by the alternative hypothesis. Rejecting H1-H8 with probability $\geq 0.90$ at an alpha level of 5%, using Holm-Bonferroni corrected p-values for multiple testing, required a sample size of approximately 500 participants per treatment. As a technical check, we ran a pilot experiment online with 60 participants recruited via Prolific. The median duration time in the first experiment was 17 minutes, and participants earned 7 pounds per hour on average. In the second experiment, the median duration time was 8.5 minutes, with average earnings of 14 pounds per hour.

**Experimental variables** We observe participants' decisions in 14 different interactions with another unknown participant from the same treatment. Our key variables of interest are the decisions participants make in the 5 two-player games, which are: the offer in UG, the minimum acceptance threshold in UG, the binary trust decision in bTG, the conditional back-transfer in case of trust in bTG, the binary cooperation decision in PD, the binary cooperation decision in SH, and the selection of the modal choice in C (earth). We generate the following normalized variables:

- Normalized offer in UG = offer/5
  Offers are usually between 1 and 5, offers larger than 5 are truncated.
- Normalized min acceptance threshold in UG = min/5
  Thresholds are usually between 1 and 5, thresholds larger than 5 are truncated.
- Normalized back-transfer in bTG = back-transfer/6
  Back-transfers are between 0 and 6 (no truncation needed).

**Data exclusion** We discard the observations of participants who report that they have been repeatedly disturbed and therefore recommend not to use their data in the analysis in an open question asked at the end of the experiment. Participants are informed that not using their data has no payoff consequences for them. We also discard the observation of participants who state in an open question at the end that they consulted ChatGPT for advice during the experiment. We also discard all decisions with unreasonable response times (either too fast or too slow). A decision is discarded if the logarithm of the time needed to make the decision is more than 3 standard deviations away from the mean.

**Research Questions and Hypotheses** We derive 8 hypotheses (H1–H8) related to 8 research questions (Q1-Q8) derived from five general assumptions about the nature of modern AI, which are outlined in the pre-analysis plan.

Q1: *How does the use of AI change social interaction?*
We compare the decisions targeted at humans to the decisions of the same participants targeted at AI in TRN.
H1: In TRN, the welfare index is smaller for decisions targeted at AI.

Q2: *Do people delegate more if delegation is opaque?*
We compare the frequency of delegation in the pooled data of TDN and TDP to the frequency of delegation in pooled data of ODN and ODP.
H2: Delegation is more frequent in ODN and ODP compared to the pooled data of TDN and TDP.

Q3: *Does the possibility of opaque delegation change social interaction?*
We compare the decisions targeted at humans in TRN to the unconditional decisions of participants in ODN (who do not know if they interact with a human or with AI).
H3: The welfare index is smaller for decisions in ODN than in TRN.

Q4: *Does delegation to AI crowd-out prosocial behavior?*
We compare the decisions targeted at AI between TDN and TRN.
H4: Prosociality index is smaller for decisions in TDN than in TRN.

Q5: *Does social interaction change less if AI is personalized?*
We compare the decisions targeted at AI between TRP and TRN.
H5: The welfare index is larger for decisions in TRP than in TRN.

Q6: *Do people delegate more to personalized AI?*
We compare the frequency of delegation in the pooled data of TDN and ODN to the frequency of delegation in the pooled data of TDP and ODP.
H6: Delegation is more frequent in pooled data of TDP and ODP compared to the pooled data of TDN and ODN.

Q7: *Does the possibility of opaque delegation to personalized AI change social interaction less?*
We compare the unconditional decisions of participants in ODP and ODN.
H7: The welfare index is larger for decisions in ODP than in ODN.

Q8: *Does delegation to personalized AI reduce the crowding-out of prosocial behavior?*
We compare the decisions targeted at AI between TDP and TDN.
H8: The prosociality index is larger for decisions in TDP than in TDN.

**Statistical analyses** We use a conditional testing procedure to correct for multiple testing. First, we test H1–H8 using Holm-Bonferroni corrected p-values, which controls for the fact that we initially perform 8 tests using an alpha level of 5%. If a test is significant, because its corrected p-value is smaller than 5%, we perform conditional tests for the same data. For hypotheses about welfare (H1, H3, H5, H7), we additionally test for differences in the predictability of behavior (frequency of modal choice in C), equality concerns (offers in UG), the kindness index, and the intentions index (4 conditional tests).

For hypotheses about prosociality (H4, H8), we test for differences in the 6 normalized decisions underlying the prosociality index: the normalized offer in the ultimatum game (offer/5), the normalized minimum acceptance threshold (min/5), the binary trust decision, the normalized back-transfer in the trust game (back-transfer/6), and the binary cooperation decision in the prisoner's dilemma and the binary cooperation decision in the stag-hunt game (6 conditional tests). For hypotheses about delegation (H2, H6), we test if the frequency of delegation differs in each game-specific situation (7 conditional tests). Simulation results suggest that the conditional testing procedure effectively restricts the familywise error rate. The simulated probability of at least one Type I error is 4.5% for the main tests and 4.3% for the conditional tests (estimated based on 10,000 simulation runs).

**Preregistration** The study was preregistered. The preregistration files of the study can be found here: `https://osf.io/fb7jd/`

# 3.5 Appendix

Table A3.1: Influence of covariates on the (loss of) welfare

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Welfare Index (base: AI) | | | Diff. in Welfare Index (Human-AI) | | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| *Decision targeted at...* | | | | | | |
| ...Human | 0.060*** | 0.060*** | 0.060*** | | | |
| | (0.004) | (0.004) | (0.004) | | | |
| ...Unknown | 0.054*** | 0.049*** | 0.046*** | | | |
| | (0.007) | (0.007) | (0.007) | | | |
| Age | | 0.0004 | 0.001** | −0.001* | −0.0003 | −0.001* |
| | | (0.0002) | (0.0002) | (0.0003) | (0.0003) | (0.0003) |
| Gender: Non-binary | | −0.037 | −0.035 | 0.101** | 0.085** | 0.081* |
| | | (0.034) | (0.034) | (0.043) | (0.042) | (0.042) |
| Gender: Woman | | −0.014** | −0.014** | −0.001 | −0.003 | −0.003 |
| | | (0.006) | (0.006) | (0.008) | (0.008) | (0.008) |
| Higher Education | | −0.008 | −0.008 | 0.001 | 0.001 | 0.0005 |
| | | (0.007) | (0.007) | (0.009) | (0.008) | (0.008) |
| Usage of ChatGPT | | 0.004 | 0.004 | 0.0001 | 0.002 | 0.002 |
| | | (0.003) | (0.003) | (0.004) | (0.004) | (0.004) |
| *Agreement to...* | | | | | | |
| ...AI difficult to predict | | −0.006* | −0.006* | | 0.017*** | 0.017*** |
| | | (0.003) | (0.003) | | (0.004) | (0.004) |
| ...AI trustworthy | | 0.015*** | 0.016*** | | −0.036*** | −0.036*** |
| | | (0.004) | (0.004) | | (0.005) | (0.005) |
| ...Equality concern with AI | | 0.055*** | 0.054*** | | −0.015*** | −0.014*** |
| | | (0.003) | (0.003) | | (0.004) | (0.004) |
| ...AI reflects human | | −0.002 | −0.002 | | −0.001 | −0.001 |
| | | (0.003) | (0.003) | | (0.003) | (0.003) |
| Distractions | | | 0.00001 | | | −0.001 |
| | | | (0.001) | | | (0.001) |
| Duration (min) | | | −0.001*** | | | 0.002*** |
| | | | (0.0004) | | | (0.0005) |
| Constant | 0.649*** | 0.472*** | 0.486*** | 0.080*** | 0.136*** | 0.119*** |
| | (0.004) | (0.019) | (0.020) | (0.016) | (0.024) | (0.025) |
| Observations | 4,808 | 4,808 | 4,808 | 1,903 | 1,903 | 1,903 |
| $R^2$ | 0.024 | 0.138 | 0.140 | 0.005 | 0.062 | 0.067 |
| Adjusted $R^2$ | 0.023 | 0.136 | 0.137 | 0.002 | 0.057 | 0.062 |

Note: OLS regressions with welfare index (baseline: decisions targeted at AI) observed in all treatments as dependent variable in models (1)–(3) and difference in welfare index (human - AI) observed in the treatments *transparent random* as dependent variable in models (4)–(6). *Usage of ChatGPT* is the numeric representation of the frequency of the use of ChatGPT. The four variables of agreement to specific statements about the AI are measured on a scale from 0–4. *Distractions* is the frequency of changing the browser window during the experiment. Robust standard errors clustered at the subject level in models (1)–(3). *p<0.1; **p<0.05; ***p<0.01

# Author Contribution

The first chapter, *Blame and Praise: Responsibility Attribution Patterns in Decision Chains*, was joint work with Deepti Bhatia, Urs Firschbacher and Jan Hausfeld. The idea for the project was conceived by Urs Fischbacher and Jan Hausfeld. The experiment was programmed by Deepti Bhatia and Jan Hausfeld before I joined the project as a master's student and continued to work on it for my Ph.D. I contributed to the final programming steps of the project. All authors except for Urs Fischbacher conducted the experiment and collected the data. Deepti Bhatia and I then analyzed the data, with helpful feedback from the other authors. The final version of the paper was written by Deepti Bhatia, Urs Fischbacher, and me, with support from Jan Hausfeld. Note that parts of this project have already been submitted in my master's thesis. For this dissertation, I revised the entire results section and performed many more sophisticated analyses.

The second chapter, *Using Mental Imagery to Foster Future-Mindedness in Long-Term Decisions*, was joint work with Baiba Renerte. She provided a preliminary version of the research idea. We elaborated on the research idea and conceptualized the project together. I programmed and conducted the laboratory and online experiment, with helpful feedback from my co-author. I also analyzed the data and wrote the paper, while Baiba Renerte supervised the process.

The third chapter *Generative AI Triggers Welfare-Reducing Decisions in Humans* was a collaboration with Fabian Dvorak, Urs Fischbacher, and Sebastian Fehrler. The initial idea was developed by Fabian Dvorak, and we both conceptualized the project. Fabian Dvorak set up the pre-registration, while Urs Fischbacher, Sebastian Fehrler and I gave feedback. I programmed and conducted the online experiment with valuable comments from my co-authors. The data analysis was done by me and the paper was written by Sebastian Fehrler, Fabian Dvorak, and me, while Urs Fischbacher supervised the process.

# Complete Bibliography


Abramoff, M. D., Whitestone, N., Patnaik, J. L., Rich, E., Ahmed, M., Husain, L., Hassan, M. Y., Tanjil, M. S. H., Weitzman, D., Dai, T., Wagner, B. D., Cherwek, D. H., Congdon, N., and Islam, K. (2023). Autonomous artificial intelligence increases real-world specialist clinic productivity in a cluster-randomized trial. *npj Digital Medicine*, 6(1):184.

Alan, S. and Ertac, S. (2018). Fostering Patience in the Classroom: Results from Randomized Educational Intervention. *Journal of Political Economy*, 126(5):1865–1911.

Albrecht, F., Kube, S., and Traxler, C. (2018). Cooperation and norm enforcement - The individual-level perspective. *Journal of Public Economics*, 165:1–16.

Andersen, S., Harrison, G. W., Lau, M. I., and Rutström, E. E. (2008). Eliciting Risk and Time Preferences. *Econometrica*, 76(3):583–618.

Andreoni, J. and Serra-Garcia, M. (2021). Time inconsistent charitable giving. *Journal of Public Economics*, 198:104391.

Ashraf, N., Bryan, G., Delfino, A., Holmes, E. A., Iacovone, L., and Pople, A. (2021). Learning to See the World's Opportunities: The Impact of Visualization on Entrepreneurial Success. In *The 2021 Behavioral Economics Annual Meeting (BEAM 2021)*.

Atance, C. M. and O'Neill, D. K. (2001). Episodic future thinking. *Trends in Cognitive Sciences*, 5(12):533–539.

Barrafrem, K. and Hausfeld, J. (2020). Tracing Risky Decisions for Oneself and Others: The Role of Intuition and Deliberation. *Journal of Economic Psychology*, 73:89–101.

Bartling, B. and Fischbacher, U. (2012). Shifting the blame: On delegation and responsibility. *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.

Barton, D., Manyika, J., Koller, T., Robert Palter, T., Godsall, J., and Zoffer, J. (2017). Measuring the economic impact of short-termism. *McKinsey Global Institute. Discussion Paper*.

Bauer, K., Liebich, L., Hinz, O., and Kosfeld, M. (2023). Decoding GPT's hidden "rationality" of cooperation. *SAFE Working Paper*, No. 401.

Baumeister, R. F., Vohs, K. D., and Oettingen, G. (2016). Pragmatic Prospection: How and Why People Think about the Future. *Review of General Psychology*, 20(1):3–16.

Bellani, L., Bledow, N., Busemeyer, M. R., and Schwerdt, G. (2021). When everyone thinks they're middle-class : (Mis-) Perceptions of inequality and why they matter for social policy. Policy Paper 6, Cluster of Excellence 'The Politics of Inequality'.

Bénabou, R. (2013). Groupthink: Collective Delusions in Organizations and Markets. *The Review of Economic Studies*, 80(2):429–462.

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1):122–142.

Bernard, T., Dercon, S., Orkin, K., Schinaia, G., and Taffesse, A. S. (2023). The Future in Mind: Aspirations and Future-Oriented Behaviour in Rural Ethiopia. *CEPR Discussion Paper*, 18492.

Besley, T. (2006). *Principled Agents?: The Political Economy of Good Government*. Oxford University Press on Demand.

Bhatia, D., Fischbacher, U., Hausfeld, J., and Stumpf, R. (2024). Blame and praise: Responsibility attribution patterns in decision chains. *Experimental Economics*, 27(3):637–663.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970):533–538.

Bieleke, M., Gollwitzer, P. M., Oettingen, G., and Fischbacher, U. (2017). Social Value Orientation Moderates the Effects of Intuition versus Reflection on Responses to Unfair Ultimatum Offers. *Journal of Behavioral Decision Making*, 30(2):569–581.

Bø, S. and Wolff, K. (2020). I Can See Clearly Now: Episodic Future Thinking and Imaginability in Perceptions of Climate-Related Risk Events. *Frontiers in Psychology*, 11:218.

Bock, O., Baetge, I., and Nicklisch, A. (2014). Hroot: Hamburg Registration and Organization Online Tool. *European Economic Review*, 71:117–120.

Bolton, G. E. and Ockenfels, A. (2000). ERC: A Theory of Equity, Reciprocity, and Competition. *The American Economic Review*, 90(1):166–193.

Bordalo, P., Burro, G., Coffman, K., Gennaioli, N., and Shleifer, A. (2022). Imagining the Future: Memory, Simulation and Beliefs about Covid. Technical Report w30353, National Bureau of Economic Research, Cambridge, MA.

Breman, A. (2011). Give more tomorrow: Two field experiments on altruism and intertemporal choice. *Journal of Public Economics*, 95(11):1349–1357.

Buckner, R. L. and Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences*, 11(2):49–57.

Busemeyer, M. R. (2024). Who cares for the future? Exploring public attitudes towards the needs of future generations in Germany. *Journal of European Public Policy*, 31(3):680–705.

Candrian, C. and Scherer, A. (2022). Rise of the machines: Delegating decisions to autonomous AI. *Computers in Human Behavior*, 134:107308.

Carton, A. M. and Lucas, B. J. (2018). How Can Leaders Overcome the Blurry Vision Bias? Identifying an Antidote to the Paradox of Vision Communication. *Academy of Management Journal*, 61(6):2106–2129.

Charness, G. (2000). Responsibility and effort in an experimental labor market. *Journal of Economic Behavior & Organization*, 42(3):375–384.

Chen, Y., Liu, T. X., Shan, Y., and Zhong, S. (2023). The emergence of economic rationality of GPT. *Proceedings of the National Academy of Sciences*, 120(51):e2316205120.

Chopra, F., Falk, A., and Graeber, T. (2024). Intertemporal Altruism. *American Economic Journal: Microeconomics*, 16(1):329–357.

Chugunova, M. and Sele, D. (2022). We and It: An interdisciplinary review of the experimental evidence on how humans interact with machines. *Journal of Behavioral and Experimental Economics*, 99:101897.

Coffman, L. C. (2011). Intermediation Reduces Punishment (and Reward). *American Economic Journal: Microeconomics*, 3(4):77–106.

Cole, S. and Kvavilashvili, L. (2021). Spontaneous and deliberate future thinking: A dual process account. *Psychological Research*, 85(2):464–479.

Dal Bó, P., Fréchette, G. R., and Kim, J. (2021). The determinants of efficient behavior in coordination games. *Games and Economic Behavior*, 130:352–368.

Dargnies, M.-P., Hakimov, R., and Kübler, D. F. (2022). Aversion to Hiring Algorithms: Transparency, Gender Profiling, and Self-Confidence. *SSRN Electronic Journal*.

Davies, R., Haldane, A. G., Nielsen, M., and Pezzini, S. (2014). Measuring the costs of short-termism. *Journal of Financial Stability*, 12:16–25.

Duch, R., Przepiorka, W., and Stevenson, R. (2015). Responsibility Attribution for Collective Decision Makers. *American Journal of Political Science*, 59(2):372–389.

Duchowski, A. T. (2017). *Eye Tracking Methodology: Theory and Practice*. Springer.

Duffy, J. and Cole, S. N. (2021). Functions of spontaneous and voluntary future thinking: Evidence from subjective ratings. *Psychological Research*, 85(4):1583–1601.

Dufwenberg, M. and Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, 47(2):268–298.

Dvorak, F., Fischbacher, U., Fehrler, S., and Stumpf, R. (2023). AI decisions in human interaction. https://osf.io/fvk2c/.

Engl, F. (2022). A Theory of Causal Responsibility Attribution. *CESifo Working Paper Series*, (9898).

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Simoudis, E., Han, J., and Fayyad, U., editors, *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, pages 226–231. AAAI Press.

Eubanks, A. D., Reece, A., Liebscher, A., Meron Ruscio, A., Baumeister, R. F., and Seligman, M. (2024). Pragmatic prospection is linked with positive life and workplace outcomes. *The Journal of Positive Psychology*, 19(3):419–429.

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., and Sunde, U. (2018). Global Evidence on Economic Preferences. *The Quarterly Journal of Economics*, 133(4):1645–1692.

Falk, A., Fehr, E., and Fischbacher, U. (2008). Testing theories of fairness-Intentions matter. *Games and Economic Behavior*, 62(1):287–303.

Fehr, E. and Charness, G. (2023). Social Preferences: Fundamental Characteristics and Economic Consequences. *CESifo Working Paper*, No. 10488.

Fehr, E. and Fischbacher, U. (2003). The nature of human altruism. *Nature*, 425(6960):785–791.

Fehr, E. and Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and human behavior*, 25(2):63–87.

Fehr, E. and Schmidt, K. M. (1999). A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics*, 114(3):817–868.

Feinberg, J. (1970). *Doing & Deserving; Essays in the Theory of Responsibility*. Princeton University Press.

Ferrera, M. (2017). Impatient politics and social investment: The EU as 'policy facilitator'. *Journal of European Public Policy*, 24(8):1233–1251.

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

Frederick, S., Loewenstein, G., and O'Donoghue, T. (2002). Time Discounting and Time Preference: A Critical Review. *Journal of Economic Literature*, 40(2):351–401.

Gaesser, B., Horn, M., and Young, L. (2015). When Can Imagining the *Self* Increase Willingness to Help *Others* ? Investigating Whether the Self-Referential Nature of Episodic Simulation Fosters Prosociality. *Social Cognition*, 33(6):562–584.

Gaesser, B., Keeler, K., and Young, L. (2018). Moral imagination: Facilitating prosocial decision-making through scene imagery and theory of mind. *Cognition*, 171:180–193.

Gaesser, B. and Schacter, D. L. (2014). Episodic simulation and episodic memory can increase intentions to help others. *Proceedings of the National Academy of Sciences*, 111(12):4415–4420.

Gaesser, B., Shimura, Y., and Cikara, M. (2020). Episodic simulation reduces intergroup bias in prosocial intentions and behavior. *Journal of Personality and Social Psychology*, 118(4):683–705.

Gerstenberg, T., Lagnado, D. A., Speekenbrink, M., and Cheung, C. (2011). Rational order effects in responsibility attributions. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society.*, 33(33):1715–1720.

Gilbert, D. T. and Wilson, T. D. (2007). Prospection: Experiencing the future. *Science*, 317(5843):1351–1354.

Gneezy, U. and Potters, J. (1997). An experiment on risk taking and evaluation periods. *The Quarterly Journal of Economics*, 112(2):631–645.

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist*, 54(7):493–503.

Gosling, S. D., Rentfrow, P. J., and Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, 37(6):504–528.

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association*, 1(1):114–125.

Greiner, B., Grünwald, P., Lindner, T., Lintner, G., and Wiernsperger, M. (2024). Incentives, Framing, and Reliance on Algorithmic Advice: An Experimental Study. *WU Vienna University of Economics and Business. Department of Strategy and Innovation Working Paper Series*, No. 01/2024.

Guo, F. (2023). GPT Agents in Game Theory Experiments. *arXiv*, 2305.05516.

Gurdal, M. Y., Miller, J. B., and Rustichini, A. (2013). Why Blame? *Journal of Political Economy*, 121(6):1205–1247.

Güth, W., Schmittberger, R., and Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, 3(4):367–388.

Hart, H. and Gardner, J. (2008). *Punishment and Responsibility: Essays in the Philosophy of Law*. Oxford University Press.

Hausfeld, J., Fischbacher, U., and Knoch, D. (2020). The value of decision-making power in social decisions. *Journal of Economic Behavior & Organization*, 177:898–912.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., and McElreath, R. (2001). In search of homo economicus: Behavioral experiments in 15 small-scale societies. *American Economic Review*, 91(2):73–78.

Hershfield, H. E. (2019). The self over time. *Current Opinion in Psychology*, 26:72–75.

Hershfield, H. E., Goldstein, D. G., Sharpe, W. F., Fox, J., Yeykelis, L., Carstensen, L. L., and Bailenson, J. N. (2011). Increasing Saving Behavior Through Age-Progressed Renderings of the Future Self. *Journal of Marketing Research*, 48:S23–S37.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Iyengar, S. (1994). *Is Anyone Responsible?: How Television Frames Political Issues*. University of Chicago Press.

Jacobs, A. M. (2016). Policy Making for the Long Term in Advanced Democracies. *Annual Review of Political Science*, 19(1):433–454.

Johnson, N. D. and Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology*, 32(5):865–889.

Jowell, R., Roberts, C., Fitzgerald, R., and Eva, G. (2007). *Measuring Attitudes Cross-Nationally: Lessons from the European Social Survey*. Sage Publications.

Just, M. A. and Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329–354.

Kawaguchi, K. (2021). When will workers follow an algorithm? A field experiment with a retail business. *Management Science*, 67(3):1670–1695.

Köbis, N., Lorenz-Spreen, P., Ajaj, T., Bonnefon, J.-F., Hertwig, R., and Rahwan, I. (2023). Artificial Intelligence can facilitate selfish decisions by altering the appearance of interaction partners. *arXiv*, 2306.04484.

Kölle, F. and Lauer, T. (2024). Understanding Cooperation in an Intertemporal Context. *Management Science*.

Kölle, F. and Wenner, L. (2023). Is Generosity Time-Inconsistent? Present Bias across Individual and Social Contexts. *The Review of Economics and Statistics*, 105(3):683–699.

Konovalov, A. and Krajbich, I. (2019). Revealed strength of preference: Inference from response times. *Judgment & Decision Making*, 14(4):381–394.

Konovalov, A. and Ruff, C. C. (2022). Enhancing models of social and strategic decision making with process tracing and neural data. *Wiley Interdisciplinary Reviews: Cognitive Science*, 13(1):e1559.

Kovarik, J. (2009). Giving it now or later: Altruism and discounting. *Economics Letters*, 102(3):152–154.

Krupka, E. L. and Weber, R. A. (2013). Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary? *Journal of the European Economic Association*, 11(3):495–524.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., and Peter Battaglia (2023). Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421.

Lave, L. B. (1962). An empirical approach to the prisoners' dilemma game. *The Quarterly Journal of Economics*, 76(3):424–436.

Leib, M., Köbis, N., Rilke, R. M., Hagens, M., and Irlenbusch, B. (2024). Corrupted by Algorithms? How AI-Generated and Human-Written Advice Shape (Dis)Honesty. *The Economic Journal*, 134(658):766–784.

Leibbrandt, A. and López Pérez, R. (2011). Individual heterogeneity in punishment and reward. *Universidad Autónoma de Madrid. Department of Economic Analysis. Working Papers in Economic Theory*, No. 2011/01.

Liberman, N. and Trope, Y. (2003). Construal Level Theory of Intertemporal Judgment and Decision. In *Time and Decision: Economic and Psychological Perspectives of Intertemporal Choice*. Russell Sage Foundation.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Costa, A. d. S., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130.

Lobeck, M. and Støstad, M. N. (2023). The Consequences of Inequality: Beliefs and Redistributive Preferences. *CESifo Working Paper Series*, No. 10710.

Logg, J. M., Minson, J. A., and Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103.

MacKenzie, M. K. (2016). Institutional Design and Sources of Short-Termism. In González-Ricoy, I. and Gosseries, A., editors, *Institutions For Future Generations*, pages 24–46. Oxford University Press.

Mathews, A., Ridgeway, V., and Holmes, E. A. (2013). Feels like the real thing: Imagery is both more realistic and emotional than verbal thought. *Cognition & Emotion*, 27(2):217–229.

Mehta, J., Starmer, C., and Sugden, R. (1994). The nature of salience: An experimental investigation of pure coordination games. *The American Economic Review*, 84(3):658–673.

Mengel, F. (2017). Risk and temptation: A Meta-study on prisoner's dilemma games. *The Economic Journal*, 128(616):3182–3209.

Monroe, A. E., Ainsworth, S. E., Vohs, K. D., and Baumeister, R. F. (2017). Fearing the Future? Future-Oriented Thought Produces Aversion to Risky Investments, Trust, and Immorality. *Social Cognition*, 35(1):66–78.

Murphy, R. O., Ackermann, K. A., and Handgraaf, M. J. (2011). Measuring Social Value Orientation. *Judgment and Decision Making*, 6(8):771–781.

Nanay, B. (2021). Mental Imagery. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2021 edition.

Noy, S. and Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192.

Oexl, R. and Grossman, Z. J. (2013). Shifting the blame to a powerless intermediary. *Experimental Economics*, 16(3):306–312.

Olesiński, B., Opala, P., Rozkrut, M., and Torój, A. (2014). Short-termism in business: Causes, mechanisms and consequences. EY Poland Report.

Oosterbeek, H., Sloof, R., and van de Kuilen, G. (2004). Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental Economics*, 7(2):171–188.

OpenAI (2023). ChatGPT Jan 30 Version.

Orkin, K., Garlick, R., Mahmud, M., Sedlmayr, R., Haushofer, J., and Dercon, S. (2023). Aspiring to a better future: Can a simple psychological intervention reduce poverty? *National Bureau of Economic Research*, No. w31735.

Pataranutaporn, P., Liu, R., Finn, E., and Maes, P. (2023). Influencing human-AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. *Nature Machine Intelligence*, 5(10):1076–1086.

Peng, C., Yang, X., Chen, A., Smith, K. E., PourNejatian, N., Costa, A. B., Martin, C., Flores, M. G., Zhang, Y., Magoc, T., Lipori, G., Mitchell, D. A., Ospina, N. S., Ahmed, M. M., Hogan, W. R., Shenkman, E. A., Guo, Y., Bian, J., and Wu, Y. (2023). A study of generative large language model for medical research and healthcare. *npj Digital Medicine*, 6(1):210.

Peters, J. and Büchel, C. (2010). Episodic Future Thinking Reduces Reward Delay Discounting through an Enhancement of Prefrontal-Mediotemporal Interactions. *Neuron*, 66(1):138–148.

Rabin, M. (1993). Incorporating Fairness into Game Theory and Economics. *The American Economic Review*, 83(5):1281–1302.

Rahal, R.-M. and Fiedler, S. (2019). Understanding cognitive and affective mechanisms in social psychology through eye-tracking. *Journal of Experimental Social Psychology*, 85:103842.

Rahwan, T., Rahwan, I., Crandall, J. W., Soroye, Z., Bonnefon, J.-F., and Ishowo-Oloko, F. (2019). Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nature Machine Intelligence*, 1(11):517–521.

Reece, A., Eatough, E., Wood, A., Wooll, M., Eubanks, A., and Liebscher, A. (2022). The Future-Minded Organization: How one mindset helps employees navigate unpredictable times. Labs Report Winter 2022, BetterUp.

Ross, K., Schumer, C., Fransen, T., Wang, S., and Elliott, C. (2021). Insights on the First 29 Long-term Climate Strategies Submitted to the United Nations Framework Convention on Climate Change. *World Recources Institute*, 21:1–14.

Ross, L. and Nisbett, R. E. (1991). *The Person and the Situation: Perspectives of Social Psychology*. New York (N.Y.) : McGraw-Hill Book Company.

Rubinstein, A. (2007). Instinctive and cognitive reasoning: A study of response times. *Economic Journal*, 117(523):1243–1259.

Ruscio, A. M., Khazanov, G. K., Reece, A., and Kellerman, G. (2023). Development and validation of the Pragmatic Prospection Scale, a measure of constructive future thinking. *Manuscript in preparation.*

Rutchick, A. M., Slepian, M. L., Reyes, M. O., Pleskus, L. N., and Hershfield, H. E. (2018). Future self-continuity is associated with improved health and increases exercise behavior. *Journal of Experimental Psychology: Applied*, 24(1):72–80.

Sampson, R. C. and Shi, Y. (2023). Are US Firms Becoming More Short-Term Oriented? Evidence of Shifting Firm Time Horizons from Implied Discount Rates, 1980-2013. *Strategic Management Journal*, 44(1):231–263.

Schacter, D. L., Benoit, R. G., and Szpunar, K. K. (2017). Episodic Future Thinking: Mechanisms and Functions. *Current opinion in behavioral sciences*, 17:41–50.

Schacter, D. L., Devitt, A. L., and Addis, D. R. (2018). Episodic future thinking and cognitive aging. In *Oxford Research Encyclopedia of Psychology*.

Seligman, M. E. P., Railton, P., Baumeister, R. F., and Sripada, C. (2013). Navigating Into the Future or Driven by the Past. *Perspectives on psychological science*, 8(2):119–41.

Seligman, M. E. P., Railton, P., Baumeister, R. F., and Sripada, C. (2016). *Homo Prospectus*. Oxford University Press.

Shaver, K. G. (1985). The Attribution of Causality. In *The Attribution of Blame*, pages 35–62. New York: Springer.

Spiliopoulos, L. and Ortmann, A. (2018). The BCD of response time analysis in experimental economics. *Experimental Economics*, 21(2):383–433.

Thielmann, I., Böhm, R., Ott, M., and Hilbig, B. E. (2021). Economic games: An introduction and guide for research. *Collabra: Psychology*, 7(1):19004.

Thompson, D. F. (2010). Representing future generations: Political presentism and democratic trusteeship. *Critical Review of International Social and Political Philosophy*, 13(1):17–37.

Tricomi, E., Rangel, A., Camerer, C. F., and O'Doherty, J. P. (2010). Neural evidence for inequality-averse social preferences. *Nature*, 463(7284):1089–1091.

Trope, Y. and Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review*, 117(2):440–463.

van Gelder, J.-L., Hershfield, H. E., and Nordgren, L. F. (2013). Vividness of the future self predicts delinquency. *Psychological Science*, 24(6):974–980.

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., and Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*, 11(1):233.

von Schenk, A., Klockmann, V., and Köbis, N. (2023). Social Preferences Toward Humans and Machines: A Systematic Experiment on the Role of Machine Payoffs. *Perspectives on Psychological Science*, 17456916231194949.

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., Anandkumar, A., Bergen, K., Gomes, C. P., Ho, S., Kohli, P., Lasenby, J., Leskovec, J., Liu, T.-Y., Manrai, A., Marks, D., Ramsundar, B., Song, L., Sun, J., Tang, J., Veličković, P., Welling, M., Zhang, L., Coley, C. W., Bengio, Y., and Zitnik, M. (2023). Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60.

Weiner, B. (1995). *Judgments of Responsibility: A Foundation for a Theory of Social Conduct*. New York: Guilford Press.

Whittingham, R. (2004). Design Errors. In *The Blame Machine: Why Human Error Causes Accidents*. Routledge.

Yan, M., Cerri, G. G., and Moraes, F. Y. (2023). ChatGPT and medicine: How AI language models are shaping the future and health related careers. *Nature Biotechnology*, 41(11):1657–1658.

Yi, R., Pickover, A., Stuppy-Sullivan, A. M., Baker, S., and Landes, R. D. (2016). Impact of episodic thinking on altruism. *Journal of Experimental Social Psychology*, 65:74–81.

Zavagnin, M., De Beni, R., Borella, E., and Carretti, B. (2016). Episodic future thinking: The role of working memory and inhibition on age-related differences. *Aging Clinical and Experimental Research*, 28(1):109–119.

Zhou, W., Lin, M., Xiao, M., and Fang, L. (2021). Exploitation and exploration: Improving search precision on E-commerce platforms. *Available at SSRN 3762144*.