UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Heidelberg University
Alfred Weber Institute of Economics

Dissertation:

# Understanding Human Behaviour Through Experiments: Perspectives on Identity & Belonging, and Methodological Reflections

A Thesis Submitted to Obtain the Degree of:
**Doctor of Philosophy (Ph.D.)**

Presented by:

Leon Houf

Supervisor: Prof. Dr. Christiane Schwieren

August, 2024

# Dissertation

Submitted to the

Alfred Weber Institute of Economics

Heidelberg University, Germany

for the degree of

Doctor of Philosophy (Ph.D.)

Presented by
Leon Houf

Supervisor
Prof. Dr. Christiane Schwieren

Subject
Understanding Human Behaviour Through Experiments:
Perspectives on Identity & Belonging,
and Methodological Reflections

Oral examination: July, 2024

# Acknowledgements

Behind this document lies a journey of $\approx 3.5$ years, and I want to thank the people who played an important role in it.

This journey began in Maastricht, and I would not have started it without the support and encouragement of Martin and Nitzan.

I am grateful to Christiane for providing the freedom and space to explore my own research agenda during my PhD.

The comprehensive and diverse body of research in this dissertation would not have been possible without my excellent co-authors. Vinicius is a great partner in crime for algorithms and AI. Guido is an opener and colleague in a completely different world (of research). While both research fields are as different as research fields could possibly be, thanks to Vinicius and Guido, I am comfortable in both at the same time. Benedikt and Rik were companions on my first large research project, from which I learned so much that I benefit from it still today. Wolf, Thomas and Jörn always provided reliable and kind support, making several projects in this dissertation possible. Souphinh, Khamthou and Saychai offered invaluable guidance and support in Laos, creating a unique learning and research opportunity for me.

I want to thank my team - Sara, Tamas, Ming, Cosima and Mai, as well as Leonie - for all the wonderful moments and friendships on this journey. To quote Leonie during data collection: "Ist ja auch schön, immer wieder neue Probleme kennen zu lernen". A motto that proved well.

For their help in administrative mysteries, I want to thank Susanne Podieh and Sina Abmaier.

Travelling was an essential part of this experience, so I want to thank everyone who made me feel welcome and at home wherever I was: Poes and Jannis in Maastricht, Britt in Utrecht, Jonas in Paris, Christina in Copenhagen, Denis, Jasmin, Yumi and family in Cyprus, and the entire Janmisay family and everyone in 'Ban Mai Sawng' for truly heartwarming hospitality in Laos.

Finally, for going with me till the very end during the challenging parts of this journey, I thank Patricia.

# Contents

# List of Figures

# List of Tables

# 1. Introduction

Understanding human behaviour has been a central quest for humanity for centuries. Current scientific endeavours are by far not the only approach to understanding humans. Myths and stories are also a central approach to building models of human behaviour that provide support and guidance in everyday life. Myths and stories have been and are being used to explain human actions and establish norms, identities, belonging, and societal expectations.

In 'Das Schloss' (The Castle) by Franz Kafka (1926), the main character *K.* travels to a village to fulfil his duty as a land surveyor. Confronted with an inaccessible bureaucracy and unwelcoming villagers, *K.* struggles to understand the norms and rules of the society. *K.* finds his expectations, hypotheses, and mental models of the society constantly falsified or ambiguously lived. Eventually, $K.'s$ desire to belong to the social circles and have the identity as land surveyor acknowledged, drives *K.* to most desperate actions. And while $K.'s$ efforts do not give any reliable or replicable results, luckily, science has given us proven methods to do so.

The social sciences are the scientific way to try to understand human behaviour and to ask questions about belonging, group identity and norms. Across all the different disciplines that are part of the social sciences, one common approach is also to build mental models of human behaviour. Economics might be the prime example of how mental models shape how a discipline asks and answers questions on human behaviour as it explicitly uses models to abstract, generalise and predict human behaviour.

Important to keep in mind, when using such models is "What we observe is not nature itself, but nature exposed to our method of questioning" (Heisenberg, 1969). One method of questioning is experimental research. Experiments serve the purpose of falsifying (or finding evidence for) hypotheses, theories and mental models. At the same time, an experimental design will always rely on assumptions, the (mental) models of the experimenter and the framework in which the experimenter conducts the experiment.

This thesis has its point of departure in experimental economic research and has two main objectives: First, six experimental studies ask and answer questions on human behaviour. The first three studies all centre around identity and belonging. The second three studies have a wide range of topics, from motivated information sampling to algorithm aversion to behaviour when defaults change. Second, all studies use different models on how to construct the experimental framework. The first three studies all use different angles to define what "identity" means in an experimental context. The second three studies all involve features that are beyond the immediate control of the researcher and thereby extend the classical experimental framework, where the experimenter controls everything besides the behaviour of the participants.

Therefore, this thesis is structured in two parts: Part 1 presents three "Experiments On Identity & Belonging," and part 2 presents three experiments with features that are "Extending The Experimental Framework".

## Part 1: Experiments On Identity & Belonging

Part 1 presents three experiments that explore questions of identity and belonging and use different models of what "identity" is. Furthermore, they reflect on the methodological approach by using interdisciplinary research, extending classical models of social science research and deriving insights on the participants of experiments.

### Chapter 2: Between Anthropology and Behavioural Economics: Interdisciplinary Complexity, Widening Moral Horizons and Belonging in a Laos Case Study

Chapter 2 starts with a lab-in-the-field experiment on identity and belonging in rural Laos that serves as an interdisciplinary reflection between behavioural economics and sociocultural anthropology. The experimental approach is a multi-level public goods game in a resettled community in rural Laos.

Through our study, we measure levels of belonging by how much participants allocate to these levels in a multi-level public good game. The first level is the individual and its' household. The second level is the social network around the individual, which we elicit through a network analysis survey. The third level is the original villages the participants belong to. The fourth level is the entire, conglomerate village that was created through the resettlement, represented by the hospital that serves the entire village. We find that women give significantly more to themselves and the people in their direct network than to the institutions representing the village than men do, which can be explained by their responsibilities for taking care of their kin.

The interdisciplinary collaboration between behavioural economics and sociocultural anthropology allows a deeper understanding of the environment in which our data were created, i.e. the data generating process. While a strict model of an experimental framework requires a clean data generating process, sometimes the "cleanness" of the data generating process is also just assumed. We explicitly observe and reflect on the data generating process, and we use the model of the 'moral horizon' as the expectations a participant commits to when taking part in such an experiment, which we describe as a "rather exceptional kind of communicative event".

The study contributes to the scientific understanding of human behaviour in several ways. To our knowledge, we are the first researchers to conduct an economic experiment in (rural) Laos. And while RCTs and field experiments are prominent tools within experimental research, we showcase how the collaboration with anthropology can be intensified to get a more thorough understanding of the data generating process behind experiments. Furthermore, we give an example of conducting interdisciplinary research between behavioural economics and anthropology that enriches the understanding of our research environment.

### Chapter 3: Going Beyond The In-/Out-Group Dichotomy: Investigating Altruism Towards Middle-Groups

Chapter 3 extends one of the most widely used models in social science research: the dichotomy of in- and out-group. This model is extremely useful to describe patterns of human behaviour and, at the same time, many situations do not fit into this model. Often, there might be a clear and bounded in-group and a clear and bounded out-group, but also groups that should not be classified as additional out-groups, but rather as middle-groups, because they share a set of identity markers with the in-group.

In our multi-lab, lab-in-zoom experiment, we create these identity markers that define the groups with identity markers from the participants' life outside of the laboratory, their choices within the laboratory and random allocation. We ask how altruistic participants act towards a middle-group. Within the task, we measure if participants break an unenforceable and nonpunishable rule in a die-rolling task for the benefit of the in-, middle- or out-group. We find that a clear hierarchy of groups exists, where the in-group is treated favourably, the middle-group fairly, and the out-group bears the loss of the in-groups gain.

This holds under two conditions: that the middle-group shares sufficient identity markers with the in-group and that the participants have gained experience with the experimental task over many rounds. This gives important insights for using models in experimental research. First, it shows that participants do think beyond a simple in-/out-group dichotomy and that situations that do not fall into this dichotomy should not be experimentally investigated as such. Second, participants need time within the study to fully comprehend more complex group settings such as the presented one. This means that the same task, with the same group setting, will yield different results when it is investigated in a short study, maybe a one-shot study, compared to when participants are exposed to the group setting for a longer time.

### Chapter 4: Measuring In-Group Favouritism on Prolific: Experimental Evidence on When Prolific Participants Show Social Preferences

Chapter 4 follows up on the model of in-group favouritism and out-group hostility. It describes a novel adaptation of the slider task to independently measure in-group favouritism and out-group hostility because these two behaviours do not necessarily need to be the inverse of each other. The experiment is conducted online on Prolific.

On Prolific, participants are recruited from the U.S. based on their support for Democrats or Republicans. A first set of participants produces a "product" which is then reviewed by a second set of participants. This performance review task effectively yields a null-finding for the tested social preferences. But when given the opportunity to altruistically give a "tip" to the workers, we observe substantial in-group favouritism.

This brings the opportunity to reflect on the operationalisation of the model of in- and out-group in this study. On the one hand, using party affiliation as the determining characteristic did not have an effect on the performance review task, but for the general altruistic option of giving points, it clearly has. From here, we can deduct a prioritisation of behaviour that Prolific participants show and will be reflected in the results. First, they try to maximise their own bonus payment. Second, they act conscientiously according to tasks and expectations within the experimental framework, giving evidence for the 'moral horizon' explained in chapter 2. And then third, when given a direct opportunity, they show social preferences. This observation of a priority in showing behaviour calls for deeper investigation, as it has implications for the study design of social preference.

## Part 2: Extending The Experimental Framework

In part 2, three studies present ways to extend the experimental framework, i.e., incorporate features into the experimental design that are beyond the immediate control of the experimenter.

### Chapter 5: Motivated Sampling Of Information

Chapter 5 introduces a study on motivated sampling of information, an intersection of information sampling and motivated reasoning which might be a potential channel leading to confirmation bias. This study explores how individuals' subjective preferences influence their information sampling behaviour when objective and subjective criteria are coupled. The subjective preferences are induced by externalities, in our study by monetary contributions to antagonising or support-worthy organisations. Thereby, these externalities are outside of the immediate experimental framework as the recipients of these monetary contributions are not involved and have no knowledge of the experiment. But the participants do have the knowledge that this externality is happening, and this knowledge about something outside of the immediate experimental framework is exactly what we as experimenters use to prompt different behavioural responses.

We present participants with a binary sampling and decision task in which they must identify the option ('computer') that uses the 'high distribution' to generate numbers. To induce subjective preferences, we vary externalities in participants' decisions by providing additional rewards or penalties for organisations the participants like or dislike. Additionally, we manipulate the type of feedback participants receive.

Our findings show that female participants sample significantly more information than male participants when faced with a negative externality or Bayesian posterior feedback. Furthermore, we identify a strong intensive margin of motivated sampling, where participants sample additionally from the option with a positive externality if they believe it is correct - a potential channel for confirmation bias.

This study contributes to the scientific understanding of human behaviour by introducing the concept of motivated sampling and demonstrating its implications for decision-making. By investigating the coupling of subjective and objective criteria, we extend the experimental framework to include externalities outside of the immediate experiment.

### Chapter 6: Trust in the Machine: How Contextual Factors and Personality Traits Shape Algorithm Aversion and Collaboration

Chapter 6 investigates algorithm aversion, i.e., how participants rely on an algorithm in a simple, repeated decision task. Here, we use treatments with the factors explainability, cost, and full task automation to benchmark their impact on participants' willingness to delegate to the algorithm. This study extends the experimental framework by placing an algorithm at the core of the experimental design, whose behaviour is outside the control of the experimenter after setting some initial parameters.

Our results show that the introduction of payment reduces delegation, whereas full automation increases it. Notably, female participants demonstrate a stronger reaction to algorithmic 'mistakes', adjusting their strategies more frequently following an error.

The importance of algorithms and AI in current times does not need extra motivation or explanation. In order to understand human behaviour in this new context, it is important to integrate algorithms and AI directly into experimental models and frameworks. This includes that sometimes, these algorithms might need to be beyond the immediate control of the experimenter. This study provides an example of how this can be done.

### Chapter 7: When To Get That Extra Paycheck? A Behavioural Evaluation Of A Default Change

Chapter 7 studies how different default options for salary payments influence the choices of civil servants in the Netherlands on when they want to receive a part of their salary. This research uses a natural experiment. Thus, the experimenters did not influence or plan the data-generating process and by construction, no experimenter-demand-effect can manifest. The researchers can only combine and analyse data sources to conduct the research.

In this natural experiment, a new policy allowed Dutch civil servants to decide when to receive their $13^{th}$ and $14^{th}$ salary payments within a year. The policy was implemented at different times across various levels of government. Some civil servants had as default an immediate, round-year payment of the salary party, while some had a default which only pays it out at the end of the year unless a civil servant takes action.

Our findings show that the type of default significantly affects adherence. Civil servants with immediate payment deviate significantly less. Additionally, employees with no experience in the old system show an even higher adherence to the immediate payout default.

Understanding human behaviour in response to policy changes is not just an academic exercise, but a crucial element of a functioning, policy-oriented democracy. This study, in the context of salary payments, demonstrates the practical application of research in shaping policies that align with human decision-making tendencies.

## Chapter 8: Outlook

Chapter 8 concludes this dissertation and gives an outlook on further research initiated in this dissertation and on new and further ways to use, adapt and extend experimental frameworks to study human behaviour.

# Part I.

# Experiments on Identity & Belonging

# 2. Between Anthropology and Behavioural Economics: Interdisciplinary Complexity, Widening Moral Horizons and Belonging in a Laos Case Study

## Authors

Leon Houf, Christiane Schwieren & Guido Sprenger [1]

## Abstract

Field studies are important methods for both behavioural economists and anthropologists, where both disciplines have distinct methodologies and expectations for results. A collaboration on eye level promises systematic research insights that neither discipline could produce on its own. This case study presents an experiment tailored for a specific place in Laos, a resettlement village composed of two ethnic groups. We find and discuss differences in the conducted network analysis and allocation in the public good game both between genders and between the two ethnic groups. Furthermore, we reflect on the experiment's positionality within the local social constructs in rural Laos. This includes Western researchers, long-term local informants, Lao researchers, and local administration officials. Our article presents insights on navigating the moral horizon of a collaborative research project between anthropology and behavioural economics and how research findings can be enriched through this process.

## Keywords

Lab-in-the-field, Belonging, Group Dynamics, Interdisciplinary Collaboration, Anthropology

## 2.1. Introduction

Belonging and identity are among the most salient topics in the social sciences in recent years and invite cooperation across disciplines. The present study approaches questions of staggered or segmentary belonging from two methodologically starkly different disciplines – behavioural economics and sociocultural anthropology. It addresses two major topics: the question with which level or context of their social system people identify with in a complex setting on the margins of market integration and the meaning of a behavioural economic experiment within this setting. Results suggest that the relative value of belonging to

[1] **Status:** This paper was presented at the CATS Forum at Heidelberg University in June 2023, at the Ecole des Hautes Études en Sciences Sociales in Paris in March 2024 and is scheduled to be presented at the Advances with Field Experiments Conferences at LSE in September 2024. Submission to an interdisciplinary journal is in preparation.

different levels of social groups may be distributed according to diverse other qualities, such as gender, and that the experiment is a form of communication that needs to be understood in relation to other systems of communication, such as state administration, kinship and ceremonial exchange. Therefore, the position of an experiment in this setting diverges from expectations regarding the standard conditions of economic experiments. This allows rethinking the question of the external validity of such experiments. It also raises the question of the boundaries of the experiment as a social event within a social system.

## 2.2. Anthropology and Economics

Since the early 2000s, there has been a growing interest in cooperation between economics and sociocultural anthropology, despite fundamentally different notions of method and empirical data (Henrich et al., 2004; Lesorogol, 2005; Stafford, 2020) and an anthropological tradition of antagonizing and politicizing this difference (Carrier, 1997). The results of this interdisciplinary research have demonstrated that the distribution of assets (such as tokens, money, gifts, . . . ) is always embedded in cultural expectations and habitus.

The exchange of commodities occurs within systems of social relationships that are ordered by specific values and ideas. Access to modern, profit-oriented markets is a crucial variable when it comes to the question how much the results of experiments converge. It seems that the closer a sociocultural setting – a "society" – is to such markets, the more similar experimental results are, while there is more diversity in results when settings are remote from them (Henrich et al., 2004).

The behavioural experiment may be a fleeting, very temporary event, but it needs to be interpreted and understood by those who relate through it, that is, the participants. These interpretations are conditioned by the ideas and values that drive social relationships in the sociocultural setting where the experiment is conducted (Henrich et al., 2004).

What an experiment and a more durable social relationship share is the fact that they both consist of communications that are produced by (a system of) communications. Insofar communications need to be understood and acted upon, they are culturally produced and mediated (see also Luhmann (1992)). Culture, here, is not taken to be a bounded entity, internally homogeneous and the reason of misunderstandings externally, but rather a process of producing communication and meaning that varies along axes of ethnicity, functional subsystem, class, nationality, gender, and other structures.

What findings by Henrich et al. (2004) suggest is that people apply the values and behavioural patterns of their everyday lives to the experimental situation likewise. Therefore, experiments need to be considered in relation to other types of relationships within a given field. This is part of the consideration of the external validity of the experiment (Naar, 2020).

However, this still does not consider that experiments themselves are events within people's social lives. Researchers take great care to separate experiments from everyday life, in order to standardise data across settings. For instance, the granting of anonymity is, apart from data protection, such a method of distinction, and so is the confinement of participants to specific spaces during the experiment. These ways of separating the experiment from other relationships, however, have very different implications in different sociocultural settings, as we argue. In many small-scale or rural societies, anonymity is a very specific and unusual condition of social interaction.

Insofar experiments are marked off from other communications and follow their own specific rules, they can be considered as small, temporary social systems within larger ones, see Luhmann (1982, 1984). As we demonstrate, these larger systems need to be mobilised in order to carry out an experiment in the first place. The social events that lead up to an experiment may differ significantly between different cultural settings. Experiments do not happen in a vacuum outside of normal social life. Even when marked as “games”, that is, special, contextually bounded communications not unlike rituals (Bateson, 1972), people get involved with them through institutions and relationships that they are already familiar with. For instance, studying at a university, and studying economics or psychology more specifically, will not only increase the likelihood of participating in such experiments. It will also endow participants with a basic understanding of the meaning and value of this – after all, rather exceptional – kind of communicative event. Many of these students even have mandatory courses on “experimental methods” and the statistical analysis of quantitative results in their curriculum. Thus, an experiment conducted with students at a university – a setting in which knowledge production is an expected activity – occupies a very different position within the relevant system of communications than an experiment conducted in a rural setting in a country where education is highly valued but access to higher education is limited, such as Laos.

In this sense, an experiment is an input in a system of communication that calls for ethnographic observation. The data produced by our research were not restricted to the quantifiable results of the experiment but also included the observation of the way the experiment found its place in the social system. It is thus to be expected that the social relationships that are activated in order to conduct the experiment in the first place would differ widely.

We suggest the term **“moral horizon”** for the social relations that need to be enacted in order for an experiment to be conducted in the first place. That is, a moral horizon is not an abstraction or a value, but a set of actors that is implied in a given action. People act because they have expectations regarding the outcome of these actions. These expectations are socially shared and commit actors – therefore, they are “moral”. Besides actors that are immediately involved, there are also distant ones who witness, create conditions or do not interfere in the action – therefore, the set of actors is a “horizon” with sometimes unclear boundaries. THE MORAL HORIZON ENCOMPASSES ALL SOCIAL RELATIONSHIPS THAT ARE NECESSARY FOR AN EXPECTED OUTCOME TO BECOME A REALITY (Sprenger, 2022, 2023). We expected to conduct a behavioural experiment that produced a certain kind of data. In order for these expectations to become a reality, we needed the commitment of certain other agents. To some of them, we engaged in immediate relationships – our long-established contacts on the village level, our collaborators from the National University of Laos (NUoL) – but we also needed more remote ones – permission from state ministries, support from the administration, funding from our home university, etc.. This expanding network of relations, some close, some remote, constituted the moral horizon of the experiment as a social event. Figure 2.1 visualises the moral horizon of our experiment.

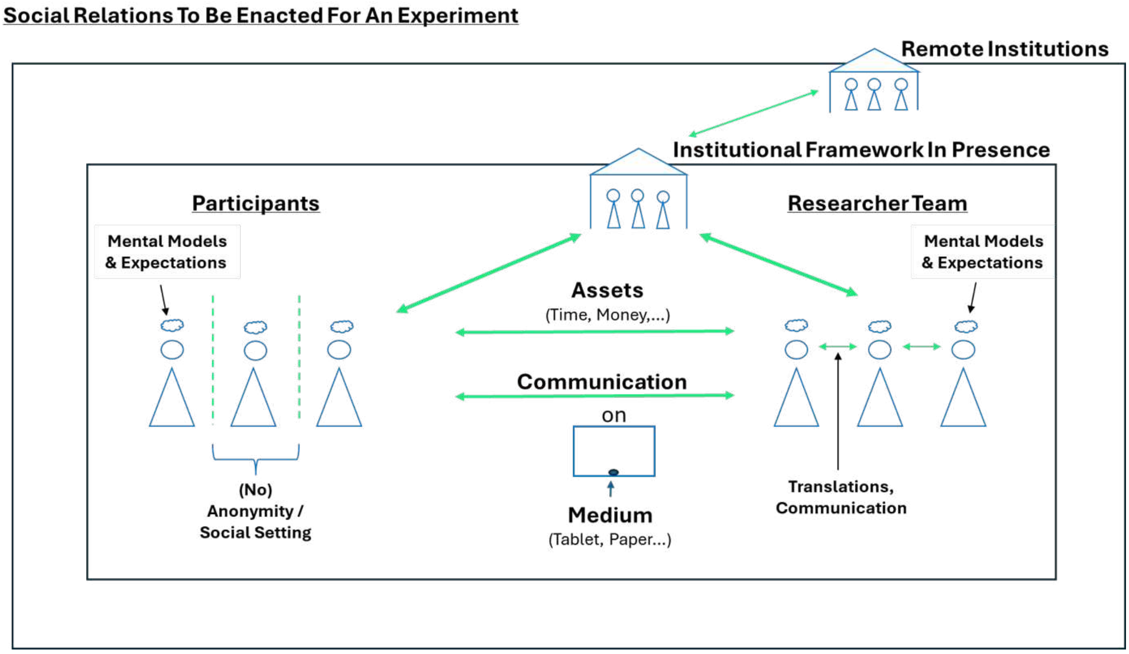

Social Relations To Be Enacted For An Experiment
Remote Institutions
Institutional Framework In Presence
Participants
Researcher Team
Mental Models & Expectations
Mental Models & Expectations
Assets
(Time, Money,...)
Communication
on
Medium
(Tablet, Paper...)
(No) Anonymity / Social Setting
Translations, Communication

Figure 2.1.: Moral horizon of our experiment.
The inner frame represents those who communicated in immediate presence and with some commitment, in time and intensity. The house on the edge of the frame represents the local administration supporting the experiment and the house outside the administration on the level of the central state.

The embeddedness of the distribution of assets becomes particularly prominent when notions of belonging are being addressed. Any social unit or category – be it a family, a place, an ethnic group or a state – is by definition a social construction that requires maintenance, and the way such units are conceived and enacted varies significantly across societies (Wimmer, 2013). These units of (potential) belonging are part of systems of social relationships. Also, we should assume that the distribution of assets is of varied importance for the integration of these entities. However, we decided that the distribution of assets is overall a good indicator for the comparative degree of belonging to such an entity. Apart from these calls for specificity, experiments, when contrasted with anthropological methods, produce comparable data from a comparatively large number of research participants. While anthropologists tend to rely on a limited number of interlocutors, with whom they collaborate intensely, and sometimes on rather general impressions, the large number of participants of experiments may reveal broader similarities among them, as well as provide data for systematic comparison (Stafford, 2020).

The behavioural economic experiment at the centre of the present study, a public good game, was thus designed to recognise the different social units that this specific setting affords. The study recognises its particular context first in the experimental design, that is, before the experiment was conducted, and second, by considering the embeddedness of the experiment in other relationships when it was actually conducted.

## 2.3. The Setting

The experiment was conducted in April 2023 in a resettlement village in Laos of ethnic minorities that was established in 2017. The Lao People's Democratic Republic, a landlocked Southeast Asian country, has been governed by a socialist party state since 1975. It officially recognises 50 ethnic groups in a population of about 7,4 million, with the largest, the ethnic Lao, making up a little more than half of it. Belonging to the Least Developed Countries (LDC), Laos's infrastructure is not well developed, and its extensive rural areas, where 80% of the population live, have uneven access to markets. In recent years, the government has made extensive efforts to graduate from the LDC status, by allowing increased foreign investment and taking up debt. Its neighbours, China, Vietnam and Thailand, have played a crucial role in this development (Vorapeth, 2024).

The site of the experiment, for which we use the pseudonym Ban Mai Sawng, is situated in Luang Nam Tha, a northern province with a high ethnic diversity. The village is composed of ethnic Khmu (also spelled Kammu or Khm'hmu') and Rmeet (Lamed), with a few Lue or ethnic Lao working there as traders, policemen, teachers or medical staff. The settlement was founded in 2017 as a resettlement village after parts of a major river valley were flooded for the reservoir of a hydroelectric dam. The dam, built as a joint venture between Lao and Chinese state-owned companies, is selling energy mostly to Thailand. Ban Mai Sawng is one of two large resettlement villages resulting from the building of this dam. The other one, pseudonymised as Ban Mai, has been studied by anthropologist Floramante Ponce (2022a, 2022b, 2023).

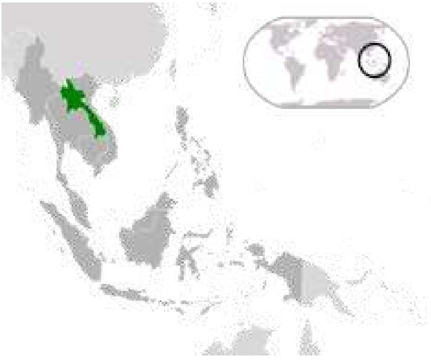


Figure 2.2.: Laos
*By Infernoapple - File:Provinces of Laos.svg, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=14976782*

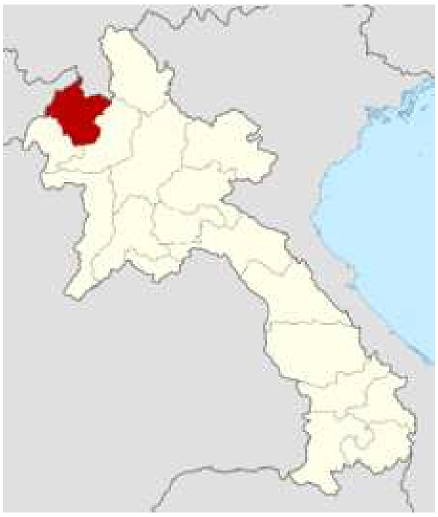


Figure 2.3.: Luang Nam Tha Region
*By ASDFGHJ - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=7350846*

This created two levels of belonging in the first place. Ban Mai Sawng is situated on land where three other Khmu villages – two of them merging in the process – had settled before, in 1985 and 2005 respectively. They were joined in 2017 by seven villages that were previously located on the banks of the river. Some of these had left their earlier mountain settings for the river bank only in the early 2000s, thus resettling twice within 15 years. These ten villages form three nucleated settlements, two of which are adjacent to each other while the third is about two kilometres away. The accumulated population of the three sites was 3676 at the time of research, with 560 households. This agglomeration contains a number of facilities that are unique, in particular a secondary school and a small hospital.

However, the planners of the resettlement had seen to keep the original villages more or less intact – former co-villagers are still neighbours now. Importantly, planners also recognized the ritual needs of the villagers. Khmu and Rmeet villages are partially defined by their respective village spirit and the ritual house in which the annual sacrifices are performed for him. The planners saw to provide most of the newly resettled villages with a ritual house and the money to perform the first ritual for the village spirit in the new location.

Local people thus perceived two levels of belonging – the former villages and the new settlement, Ban Mai Sawng. This differentiates a debate about Southeast Asian – and other – villages. There has been a longstanding assumption, shared by both scholars and policy makers, that the village is the prime site of social identification and solidarity in Southeast Asia. However, more recent research has been wary of this idea, showing that village solidarity, while important, is, on many occasions, second to the solidarity of households who pursue other aims. Village solidarity as the singular focus of belonging and economic or political action thus appears as a modernist fantasy about rural lives that overemphasises the contexts in which such solidarity is expressed (Evans, 1995; High, 2014; Kemp, 1988; Sprenger, 2021). It should be assumed that this significantly affects the distribution of assets.

Two more levels are also discernible, the level of the household and of kinship and friendship networks. Households in this context are defined as houses occupied by extended families, usually a married couple with unmarried children, but often (in about 50%

### Local Kinship Networks

Local kinship networks are diffuse, but usually contain the households of married siblings, children and parents. Among both Rmeet and Khmu, special importance is given to the relation between son-in-laws' and wife's parents' households. Both ethnicities are patrilineal, and in both, the house from which a house's wife originates from is considered superior in terms of respect owed. Wife-giving households are considered the source of fertility of household and rice fields. Therefore, wife-takers transfer gifts to their wife-givers in the context of numerous rituals, including agricultural rituals (sowing, harvesting), the life-cycle (birth, housebuilding, death) and crisis (illness) (Lindell, Samuelsson, & Tayanin, 1979; Sprenger, 2006b; Stolz, 2021). Wife-giving households receive large gifts – mostly animals and money – during weddings and funerals as well as small sacrifices or containers (pots, jars etc.) and occasional gifts of different kinds during other occasions. Wife-givers reciprocate mostly in ritual services (Sprenger, 2006a; Stolz, 2020). While immediate in-laws are preferably addressed in these rituals, the respective categories of wife-givers and wife-takers can be expanded quite significantly, according to spatial proximity or preferred alliances. While the research design did not include this complex differentiation of kin types, future research would find important dimensions of distribution of assets here. What we did see, however, was a recognition of the outmovement of daughters that the patrilineal arrangements brought about (see below).

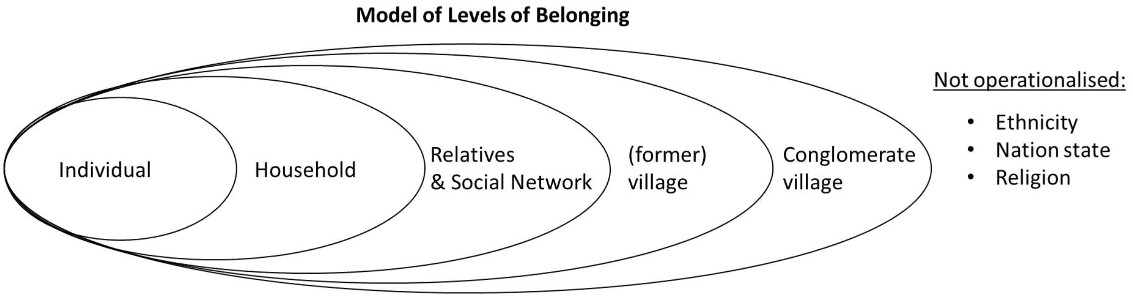


Figure 2.4.: Levels of Belonging in our Study

of cases) supplemented with members of the previous generation, usually the husband's parents. These units are called nya among Rmeet and kaang among Khmu, both terms referring to the building and the family likewise (Sprenger, 2006b; Stolz, 2021; Svantesson et al., 2014).

Additional levels of belonging and identification in this context are ethnicity and nation state. Numerous scholars have shown that ethnicity is fluid and contingent in this region (Bouté, 2009; Evrard, 2007). More recently, it seems that administrative procedures of normalisation, such as censuses and identity cards that require citizens to specify their ethnic identity have contributed to reducing such ambiguities. Regarding ethnic affiliation, the majority of inhabitants of Ban Mai Sawng are ethnic Khmu (exact numbers are unavailable), while a significant minority are Rmeet. At about 710,000 in 2015, Khmu are the largest ethnic minority in Laos and form the majority in some districts of Luang Nam Tha province. The Khmu language is thus a second lingua franca in the province, besides Lao. Rmeet numbered 22,000 in 2015. Both ethnicities speak Mon-Khmer languages, have a tradition of dry rice cultivation on shifting fields in mountainous regions and address spirits of ancestors and the locality in their rituals. They consider each other as closely related in both language, economy and culture, in comparison to the majority ethnic Lao who are associated with Tai-Kadai languages, wet rice cultivation, historical states in the lowlands, and Buddhism. Figure 2.4 visualises a model of levels of belonging in our study.

**(Non-)Buddhists and Asset Distribution**

The difference between Buddhists and non-Buddhists has an influence on the distribution of assets. Without Buddhism, there is no alms giving to monks or support of temples, but the sacrifice of domestic animals to a variety of spirits, with small sacrifices in the annual cycle and larger ones on the occasion of illness or misfortune. For both Khmu and Rmeet, the annual village ritual, in which a pig, a cow or a buffalo are killed (varying by village and year), is an occasion for shared expenditure and feasting. It also supports identification with the village community. If the experiment had covered both Buddhists and non-Buddhists, the data collected and the design would have to recognise this difference. The same goes for the level of the nation-state – all participants were Laotian citizens - , although certain ideas connected to the nation-state and its representatives do play a role in the experiment.

In regard to the market integration of Ban Mai Sawng, access to the nearest market is about one hour by car or motorcycle, with only a limited number of houses having access to the former. Ban Mai Sawng has a number of small shops selling household necessities, processed food and clothing, and a larger Chinese-owned store that also offers building material. Farmers also sell produce, especially pigs, zebu cows, and buffaloes. However, much of local food subsistence is produced by farming.

Many men and women have gathered experience as temporary labourers in towns, both in Laos and Thailand. It seems that a majority of households has people of the young generation working outside the settlement, although often temporarily and on menial jobs. Regular and important labour migration from this region can be dated back at least to the late 19th century (Amnuayvit, 2017), and for the new settlers in Ban Mai Sawng, it has increased significantly – mostly due to the fact that the three villages that have moved to the site in the preceding decades claim all the surrounding land, leaving little to cultivate for the 2017 newcomers within short distance.

However, people in rural Laos make fairly clear distinctions between the sociality of the market and life in the village, thereby orienting themselves toward different values in the two contexts. In addition, there are a number of people with comparatively little experience in market relations. This concerns elderly minority women, who have much less experience with travelling and labour migration, but also with selling produce.[2] For reasons explained below, these people were present in our sample to a disproportionate degree. During Sprenger's earlier fieldwork, Rmeet villagers in general stressed that they do not know how to trade and make money, in comparison with the ethnic majority. This self-deprecating attitude was formulated as part of their ethnic identity, that, in reverse, stressed mutual care and love. Despite the growing presence of the market, local livelihood is still strongly oriented towards subsistence in the resettled village. We are not aware of any Khmu or Rmeet household not growing its own rice and raising its own buffaloes, zebu cows, pigs or chicken. In addition, vegetables are grown in garden plots and collected in the wild. It seems that only a few food items such as processed food (dried noodles), eggs or salt are bought, and mostly not on an everyday basis.

As hinted at above, there was considerable pressure on the economy of the resettled newcomers, as the land surrounding the village cluster was used by the two (originally three) villages that had settled there in the 1980s and early 2000s. As there is little flat land in the region and a low population density, Khmu and Rmeet traditionally are shifting cultivators of dry rice. This requires large areas of land that lies fallow in between cultivation periods. Therefore, most newcomers had to travel several hours to reach their fields or look for paid work outside of the village. As explained below, this had an important influence on the recruitment of participants.

[2] This is relative to men's experience. Many younger women and some elderly do have experience in labour migration, but overall less than men.

## 2.4. Methodological Challenges

Our method combined ethnography with a behavioural experiment. Guido Sprenger has conducted anthropological research among Rmeet since 2000, intermittently for about two years, and has known some of the households that have been moved to Ban Mai Sawng since that time. He has focused on the distribution of assets through ritualized exchanges. However, he is less familiar with household economy in the resettled village than in previous field sites. Conversations between him and Rmeet informants were conducted in Rmeet, with other informants in Lao with the help of a Rmeet or Lao interpreter.

Sprenger prepared the experiment on site with conversations to the local village chiefs. A week later Leon Houf arrived together with three colleagues from the National University of Laos. They had translated the text of the experiment to Lao, interpreted in the village and supported villagers with explanations of the experiment.

The experiment included three sections: demographic information, network of respondents, and a public good game in which participants have the opportunity to distribute assets endowed to them through the game framework to either themselves or to a range of other people. The assets allocated to the most public level are doubled by the researchers, with the idea that by giving on this level, the returns will benefit everybody.

The first section of the questionnaire asked for the following information: address (section of resettlement village), role in household, gender (by identity document and by self-identification), language capability, profession, income type, ethnicity, education, original village, annual income, household size and composition.

The second section of the questionnaire aimed at asking for the network of the respondents. Participants could fill in a variety of answers to questions: With which other households do you work your fields together? Which households have you lent money to or from? With which households do you share food? With whom do you have private contacts in general? Which households do you help in ritual? For these contacts, participants could rank up to five oftheir most important contacts and provide a name and a drop-down category of the social relation to that person (relative, friend, village head, . . . ).

In the third section of the experiment, people could distribute 50 tokens, one token representing 1,000 kip. As the hourly wage of several people interviewed, who are working in Thailand, corresponded to about 35,000 kip per hour, 50,000 seemed a proper amount. These 50,000 could be distributed to one's own household, to one of the households of the network mentioned in the second section, to the rituals of the village spirit (representing the level of integration of the original village) or to the hospital (representing the common good for the resettlement village). Thereby, the operationalisation of the common good was a real entity in the life of the participants. After the end of the data collection, the researchers prepared envelopes containing the money that people had distributed, except for the hospital, which could unequivocally be separated from the other amounts. Anything given to the hospital was in the end doubled (this had been announced) and handed over to the hospital after all experimental sessions concluded, so that there was an incentive for giving to that level that was socially the most remote and general.

The experiment was thus specifically designed for this particular village, even though many of its features could be transferred to other, comparable fields. This also came with methodological specifics regarding the broader circumstances it was embedded in. The moral horizon of this experiment, the condition of its realisation, diverged from the standard settings of such experiments in universities, and this had an immediate impact on the way it was conducted. From a strict economics point of view, one may argue that this leads to a contamination of the resulting data. However, as explained above, we consider experimental games as communicative acts that are embedded within other communicative processes. It is therefore important to recount the circumstances of how the data were created in more detail to understand the insights the data can give us.

### 2.4.1. Money as a Means of Exchange

While using tokens of money in experimental games is standard, it amounted to a conscious reflection and decision in the present case. In the uplands of Luang Nam Tha province, several currencies are circulating for various purposes (Alary, 2021; Sprenger, 2007). Besides the national currency, the kip – highly inflationary at the time of fieldwork – Thai baht represents a more highly valued and more stable currency. Both are used for market as well as for ritual exchange. Besides that, silver coins from the colonial era (Piastres de Commerce) mostly serve as ritual exchange objects that can be bought and sold on the market and thus work as limited-use money. In addition, livestock or rice may be used for transfers, especially among uplanders, albeit in a restricted way. Our choice for using the national currency thus positioned us as representatives of a specific sphere of exchange, that represents the state. The kip is indeed the most common currency to date, even though people are aware of its alternatives. Using the kip also reinforced our association with the state that also conditioned our arrival, by official research permission and accompaniment of researchers from the capital city. Thus, while using the kip – or state-issued money, for that matter – for our exchanges in the experiment seemed the most plausible option, it was neither self-explanatory nor socially neutral.

### 2.4.2. Recruitment of Participants

As a first step, Sprenger asked his long-established host family, in particular his host brother, a former schoolteacher well-known in the area, for help. They visited the (elected and government-approved) headmen (*nai ban*) of the three new subsections of Ban Mai Sawng and the two older villages and explained the game to them. Each of the nine (old and new) resettled villages had kept its own headman, but in each of the three large sections, one of them was assigned the position of the main headman. This introduction to the headmen was necessary, as foreigners in rural Laos are a rare sight, and representatives of the administration such as headmen or the police need to be informed about their doings. The headmen then decided that the experiment would be conducted during three days in five different sites, according to the segments of the settlement: the old and the new part of section 1 of the conglomerate village, section 2, and the old and the new part of section 3. For the experiment, they used the English loanword to Lao, *gem* (game), marking it as a new and foreign activity.

Several factors influenced the selection of participants. As mentioned above, the economic situation of the resettlement village caused newcomers to either work outside of the area, in cities in Laos or Thailand, or to grow rice on remote fields, leading to long absences during the daytime or even of several days. Both caused a lack of people able to perform hard manual labour in the fields, in particular men, among participants, and a relatively larger number of the elderly – thus almost reversing the demographics of experiments at universities. However, in one of the old Khmu villages, the demography represented in the experiment was much more even and more representative of the actual population. This was due to the fact that there had been a death in the village two days before and a ritual prescription kept people from leaving.

We ended up with 329 participants, of whom 233 identified as Khmu, 69 as Rmeet, 15 as Lao and 4 as Lue, the rest not giving ethnic identities. The great majority identified as farmers (304). 176 identified as male, 133 as female, while 20 gave no clear answer. Regarding income, almost everybody reported that income was derived from the sale of livestock, some also from the sale of vegetables or forest produce. 93 participants reported that they did not attend school at all (or did not finish it), and 153 that they finished primary school. Only 53 and 24 respectively, reported that they had finished Lower and Upper Secondary School.

| **Education** | **Female** | **Male** |
|---|---:|---:|
| Did not attend school | 62 | 29 |
| Primary School | 53 | 94 |
| Lower Secondary School | 8 | 43 |
| Upper Secondary School | 7 | 16 |

Table 2.1.: Education levels by gender (N)

### 2.4.3. The Experiment Within the State-village Relationship

Given the necessity to inform the headmen and administration about our plans, the recruitment of participants – that is, the establishment of a short-term relationship with them – was embedded in longer-lasting relationships that followed local requirements. The actual experiment showed that it was not just a matter of a relationship between researchers – usually representatives of universities – and participants, which are often students. Rather, the state administration was also present, giving the event the air of an official occasion. A narrative of the events will demonstrate this.

On the first day of the experiment, the village headmen of four of the resettled villages, forming Section 2 of Ban Mai Sawng, informed villagers via a public announcement system (loudspeakers on poles) about the experiment. About 48 people gathered in a class room of the primary school of the section. The four headmen of the former villages and their deputies, eight men in all, sat at a long table lining the side of the room ('B' in figure 2.5). From the teacher's desk in front, one of our partners from the National University of Laos explained the experiment (to the left in the picture), and a representative of the municipal

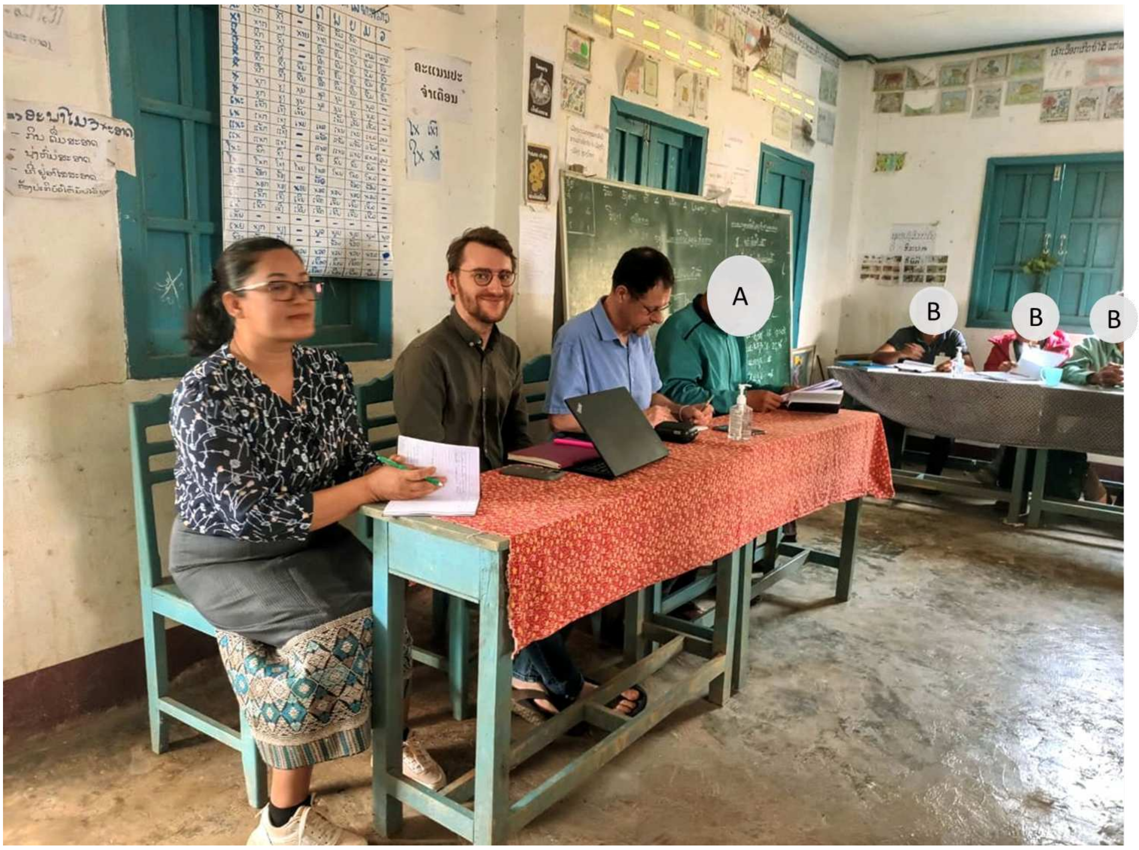


Figure 2.5.: First Experiment
*Photo taken by Khamtou Kanyavong*

administration (person ‘A’) urged the participants to be honest with their entries, while the headmen were content to quietly look on.

Especially the representative of the municipality performed in the Lao state’s function as leader and educator of the people, a typical image projected by socialist states, see Singh (2014). This also reinforced the difference between the researchers from outside – from the capital city and Germany – and the villagers. Especially the demand to be honest implied that the latter are in need of moral advice.

The relationship between villagers and state is quite specific and ambiguous in Laos, see High (2014); High and Petit (2013). The word pasason at once means “peasant” and “the people” – the latter in the sense of the sovereign of a Socialist state, as in “Lao People’s Democratic Republic”, the official state name. However, when asked about his or her profession, a peasant would reply “pasason”, in contrast to any other profession such as teacher or official. In that sense, “pasason” is the unmarked category in the field of making of living, the common background from which other professions are distinct from. The Lao party-state and its agents represent that shared background, but also hierarchise it, by leading “the people”. Sprenger has observed such advice giving on multiple occasions in various parts of Laos.

Our experiment thus was embedded in the reproduction of the proper, valorised hierarchy between the state administration and the peasantry. It drew upon relationships that are not accounted for in the experiment’s design but also gained legitimacy in the eyes of the participants.

### 2.4.4. Privacy

The safeguarding of privacy is a crucial element in economic experimenting as well as network analysis, for both data protection and the focus on individual decision making. The standard for this is a specific procedure.

1. **Private actions and exchange:** Participants would fill in the questionnaires and distribute tokens in a separated space, such as a cubicle, on a tablet computer. The money they receive would come in an envelope.
2. **Unique data identification:** Individuals are identified with a unique data identifier. In any token exchange between participants or network identification this identifier needs to work across all of the participants. In a computerised lab experiment this could be a number (e.g., participant 1, 2, 3), in a field setting, such as in our study, this could be the personal name or address. This personally identifiable information is then replaced by a simple code, when data are being processed. Through this procedure the dataset could be published with the codes only, while it allows participants to distribute tokens to other participants and also for the researchers to allocate money to participants – money for the tokens they took for themselves and received from others.

However, such a procedure, while carefully geared to ascertain anonymity, comes with a number of social assumptions, beginning with the use of (tablet) computers. Any technology relies on assumptions about its users, and even technology that is flagged as "user-friendly" or "self-explanatory" is not universally applicable. One obvious point here is literacy. The great majority of villagers were fluent in the national language, Lao, while there is no common transcription for Khmu or Rmeet. We thus decided for Lao as the language of the experiment. Previous research had suggested that a growing number of people in rural Laos has learned to handle cell phones, including the parent and grandparent generation, as they were keeping in touch with relatives working outside of the village. However, this turned out to be true only to a limited degree. At least, the abilities of most older villagers were insufficient to handle the tablets which were used for the experiment. Some claimed not to be able to read – this in particular was true for women beyond 50. Others said they could read but not the small type on the screen. Still others had difficulties operating the touchscreen. A few elderly women did not master the Lao language at all.

Arguably, these are factors not unusual in rural settings. What we want to point out here, however, is the telling way this problem was solved in all of the five occasions when the experiment was conducted. The issue was obviously considered a communal matter. Anonymity was dropped. Young people able to handle cell phones quickly picked up the skills to handle tablets as well. The younger participants helped the older to fill in the personal data and explained again the distribution of points. The researchers from NUoL actively and patiently advised participants, sometimes doing so in one-on-one sessions, sometimes with three or four people simultaneously. Little crowds gathered around the tablets, watching people filling in the questionnaire and listening to explanations.

While being alone and doing things by oneself is not entirely ruled out in rural Laos, it is a somewhat exceptional situation, and problems are usually solved by debate and cooperation (Stobbe, 2015). Solving a problem or making a decision in privacy, as the default mode of experimental games suggests, is thus not culturally neutral, but relates differently to social life outside of the experiment. While making decisions in privacy is the default mode in Germany and a number of Western-modern countries, making decisions while in contact with others – spouses, relatives, other villagers – seems to be the default mode in rural Laos. Even during school exams, as Huijsmans and Piti (2020) observe, teachers sometimes help minority students and allow them to copy each other's answers.[3] Therefore, counterintuitively, the seeming contamination of data through non-anonymity in our research is closer to isolated decision making inside and outside the experiments conducted at Western universities. Just as Western university students and people more generally are accustomed to solve problems in privacy, both within and outside an experiment, peasants in Laos are used to do this via debate and cooperation, apparently in both contexts. It is difficult to measure how this relates to the external validity of the experiment in these two respective settings.

### 2.4.5. Personal Data

In rural contexts with comparatively low educational standards, the acquisition of standardised data may be a problem. Once again, this should not simply be considered a deficit, but an indicator of differences in the distribution of knowledge. Education – meaning state-led education – standardises knowledge and makes people more aware of administrative needs. Therefore, people are getting accustomed to position themselves in terms of data that are relevant to the state, such as annual income, postal address, age or date of birth. People who are less close to the state may identify with such items of information much less, leading to gaps in the records or mere fictions designed to satisfy administrative needs. The main problem in studying a social network, as it turned out, was the difficulty to find a unique identifier for each person that was commonly known to everybody else in their network. Names are multiple. The official names on identity cards are usually structured as Lao names, with three- to four-syllable names for first and family names. However, in everyday conduct, teknonyms that refer to offspring – such as "Mother of X" or "Grandfather of Y" – are much more common, with the name of children usually reduced to a single syllable. These one-syllable names may also be used in everyday conduct. Not everybody knows the full official names within one's network, while teknonyms and one-syllable names are fairly conventional and may refer to several people. For the data base, this may result in several names for the same person, and at the same time in the same name for several individuals. This makes it unfeasible to produce results of a computational network analysis that could reliably represent the underlying

[3] As Huijsmans and Piti note, teachers were fully aware that their behavior was irregular by "developed" standards, but they valued the shared national effort to master development goals even higher – an indicator that communal problem solving is the default.

social network, at least with standard methods. In addition, while each house has an address, this is not general knowledge either. Even people living in a given house, I found, often need to check their official address in their documents. Working with photographs or maps proved unworkable under the given conditions. Thus, we were unable to find an easy-to-use, universal unique identifier that was shared by everybody in a given network, within the brevity of time. This largely impeded our efforts to link the datasets into an overarching network. What we still had, however, were data on the way people preferred to distribute their assets according to the levels of identification (see figure 2.4).

In the present case, these issues once again showed a gap between young and middle-aged people on the one hand and the elderly on the other. Elderly people sometimes did not know their age or dates of birth. The dates given in their identity cards thus sometimes were estimates or fictions. Some knew their year of birth in the Rmeet or Khmu calendar – a sixty-year cycle that does not add up years infinitely, as the Christian calendar does, but still easily translatable into it. Others knew their year of birth both in the modern and the Rmeet calendar.

More importantly, numerous people had difficulties estimating their annual income, as money plays an important, although restricted role in their household economies. As mentioned above, a large part of food is raised, grown or collected. This is certainly true for the older Khmu villages. Work sharing, that is, help in sowing, harvesting or other chores, is accounted for in man-days and balanced over time. Thus, quite a large part of economic activity is not accounted for in monetary terms.

Other parts are accounted for in terms of money. This applies to produce sold to traders or on markets, especially animals such as chicken, pigs, zebu cows, or buffaloes. Fish are raised in privately owned fish pools for sale as well. However, income from such sources is irregular, and the people Sprenger talked to had difficulties estimating an annual income from it. However, when accounting for money income, most mentioned the sale of livestock, and sometimes vegetables and forest produce. Five reported trade as their most important source of income, and these probably included the four people who reported traders as their occupation. Four mentioned salary as income, which may match the three government officials taking part.

The other important monetary income comes from wage labour – although, surprisingly, this does not figure in the data we collected. Types of labour include construction work on private construction projects (not formal companies, for men), plantations (for men and women) and, to a lesser degree, domestic work (for women) inside Laos. Opportunities for wage labour are somewhat more diverse in Thailand, where they are offered in clothing shops, restaurants and food processing factories (mostly women) as well as rice mills, gas stations and multiple other businesses. It is unclear why this did not figure more prominently in the data. First, it is possible that respondents only accounted for their own sources of income, and not that of other household members. Thus, the fact that we collected data in the village may have biased our data in this respect. It is also possible that in this context, people primarily identified as peasants and did not consider their occasional, irregular labour trips as part of their reliable income. This may have been

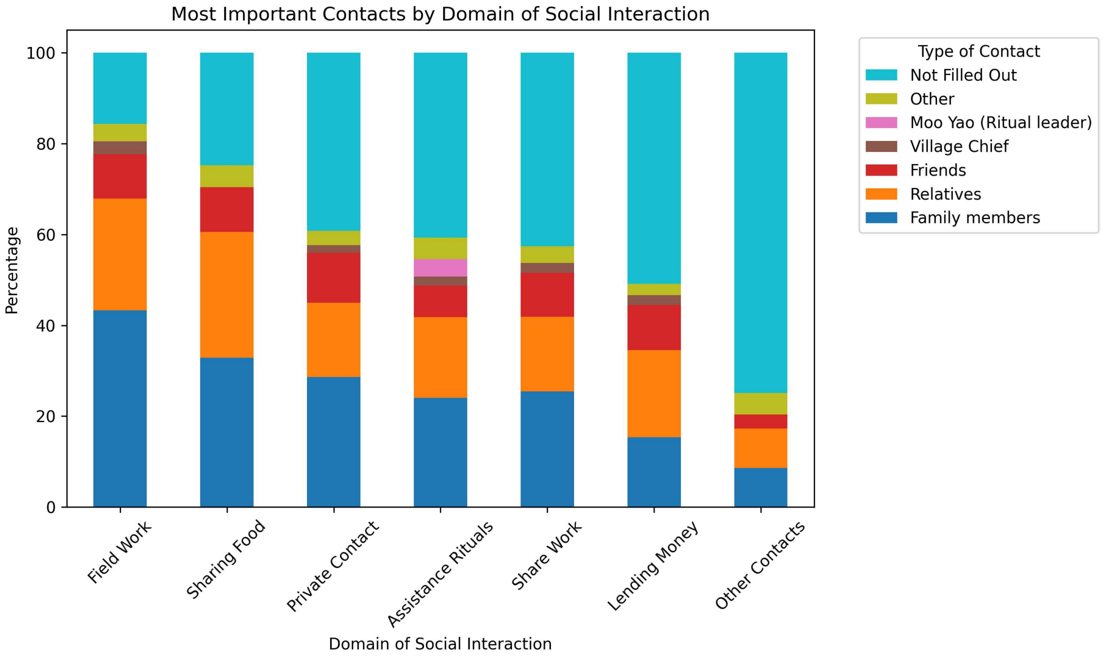


Figure 2.6.: Social Contacts

further influenced by the presence of the researchers from the National University of Laos. The administration is aware that quite a number of these labour trips are illegal, but does not, as far as we know, operate serious concerted action against them. However, we are not sure in how far people are aware of this leniency; and there is also a potential issue with tax evasion, which is likely to be an important factor. All of these – still somewhat hypothetical – possibilities may have limited a full account of the role of household income from labour migration.

## 2.5. Outcomes: Social Contacts

In section two of the survey, participants could name and categorise their most important social contacts across several domains. Figure 2.6 shows which type of contact (e.g., Family members, relatives, friends, the village chief, . . . ) participants named in their top three contacts per each domain of social interaction (e.g., fieldwork, sharing food. . . ).

The graph includes as top category in each bar, "Not Filled Out" by the participant, as a graphical representation in which domains participants named the most and the least other people. In the domain of fieldwork, participants filled out 84% of all possible top 3 positions in sharing food 75% and in lending money only 49%.

Over all domains, there is a clear order of which types of contact are named the most. First, family members have the most mentions (except for lending money), second are further relatives and third are friends.

## 2.6. Outcomes: Principles of Distribution

In section three of the survey, participants allocated tokens in a public good game that were transformed into money by the researchers at the end of the experiment.

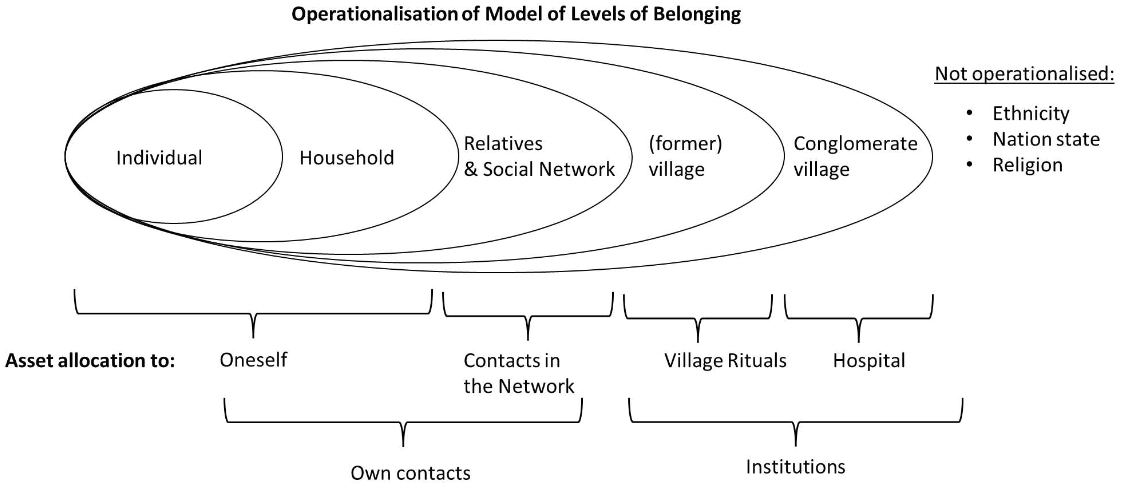


Figure 2.7.: Operationalisation Levels of Belonging

Figure 2.7 shows the operationalisation of a model of different levels of belonging. The first two levels, "Individual" and own "Household", are covered by the option to give tokens to "Oneself". Relatives and other contacts in the social network are covered by "Contacts in Network". Together, these two options will be summarised as "Own contacts". The sub-village level (i.e., former village unity) is covered by the option to contribute tokens to village rituals. The conglomerate village is covered by the hospital, an institution that serves every sub-village. These two options will be summarised as "Institutions".

Table 2.2 shows the average outcomes of distributions in the public good game. The columns represent the different options and their representation in our model of levels of belonging explained above. The first row reports the results of all participants, and the subsequent rows split these observations by gender, ethnicity, and age.

| **Participants** | | **N** | **Own Contacts** | | | **Institutions** | | |
|---|---|---|---|---|---|---|---|---|
| | | | Oneself | Network | **Total** | Rituals | Hospital | **Total** |
| All | | 329 | 16.7 | 18.4 | **35.1** | 9.2 | 5.7 | **14.9** |
| Gender | female | 133 | 18.3 | 22.1 | **40.3** | 5.8 | 3.9 | **9.7** |
| | male | 175 | 15.3 | 16.3 | **31.7** | 11.6 | 6.8 | **18.3** |
| Ethnicity | Khmu | 233 | 16.7 | 16.3 | **33** | 10.8 | 6.2 | **17** |
| | Rmeet | 69 | 16.9 | 24.3 | **41.2** | 4.6 | 4.1 | **8.8** |
| | Lao | 15 | 18.2 | 19.3 | **37.5** | 7.6 | 4.9 | **12.5** |
| Age | under 30 | 75 | 15.2 | 16.4 | **31.6** | 12.4 | 6 | **18.4** |
| | 30-60 | 160 | 17.3 | 19.7 | **37** | 7.8 | 5.2 | **13** |
| | above 60 | 67 | 16.8 | 20.9 | **37.7** | 7.4 | 4.9 | **12.3** |

Table 2.2.: Summary of Distribution
*(Only subcategories with at least 15 observations.)*

While time in the field did not allow for intensive research on the principles by which people distributed points in the game, we collected a number of statements and brief interviews on this issue that we integrate with a statistical analysis of the outcomes of distribution.

Most people gave similar amounts of tokens to their own household and to their network. On average, participants allocated 16.7 out of the 50 tokens (i.e., 16,700 Lao Kip) to their own household and 18.4 tokens to households they mentioned in the network questionnaire. This difference between allocation to the own household and people in the network is not statistically different from each other ($p=0.0799$ with a Wilcoxon Signed-Rank Test).

At the broader level, we compare the amount of tokens participants gave to their own contacts with the amount of tokens they gave to the institutions. On average, all participants gave more to their own contacts (35.1) than to the institutions (14.9). This is highly statistically significant ($p<0.001^{***}$ Wilcoxon Signed-Rank Test).

Here, especially women gave more to their own contacts than men. Women gave, on average, 40.3 tokens to themselves and their own network, statistically significantly more than men, who gave 31.7 tokens ($p< 0.001^{***}$ with a Mann-Whitney U Test). There are several ways of interpreting this finding. Household decisions are, according to interlocutors, made by a married couple together (see also Ornetsmüller, Castella, and Verburg (2018)). While women usually have more control of the household budget, with the increase of labour options in faraway places, men have become more prominent (Ireson-Doolittle, 2004). The political system also favours men, including the headman, an institution that emerged in the late 19th century, when villages in this part of Laos were increasingly integrated into trade and administrative networks (Izikowitz, 1979). Thus, men have become representatives for the world beyond the village. Therefore, they may identify more with the largest, administratively acknowledged entity they live in. This was indirectly corroborated by the elaborate excuse Sprenger received from one of the village headmen, who explained how he regretted not giving anything to the hospital despite knowing how important the institution was. In this case, women would perhaps identify more with their networks of kinship.

Also, the Rmeet gave, on average, 41.2 to their own contacts, compared to 33 of the participants who identified as Khmu and 37.5 who identified as Lao. Taking care of multiple hypothesis testing, the difference between Khmu and Rmeet is statistically significant ($p<0.001^{***}$ Mann-Whitney U Test). We want to interpret this result with care, as there are multiple potential explanations and thus, we want to rather treat it as a hypothesis for further research.

A possibility is the fact that the Rmeet consider themselves as a minority in the nation and the province, while Khmu have significantly larger numbers and also significant influence in the provincial administration. Some Rmeet stress a desire for representation and acknowledgement on the national level (Sprenger, 2017), in a way that parallels the Khmu. In that case, a reluctance to keep assets within one's network would mirror a sense of disidentification with larger groups, especially those created by state actors, such as the resettlement village. There is an additional possibility that our data on household income does not capture significant wealth differences on average between Khmu and Rmeet. With larger networks reaching well into the administration, the Khmu may have better conditions to become well-off and thus a stronger tendency to invest into larger social entities (Sevenig, 2015). However, additional research needs to verify these hypotheses.

Within the different age groups, young people gave the most to institutions (18.4 tokens), compared to people between 30-60 (13.0) and people above 60 (12.3). Correcting for multiple hypothesis testing, this is not statistically significant. We want to treat it as a potential hypothesis for further research, with two potential explanations. On the one hand, young people have fewer children to whom they could give money. On the other hand, younger people spend longer times at school than previous generations, thus being more exposed to the rhetoric of the Lao socialist party-state (Huijsmans & Piti, 2020). This rhetoric strongly emphasises the 'solidarity' (samakhi) of all ethnic groups and levels of Lao society, thus encouraging giving on communal levels and among (relative) strangers.

Whether Khmu resettled to the field site decades ago or just recently does not matter in how much they give to their own contacts. Khmu that resettled long ago (N=85) gave an average of 34.0 tokens to their own contacts, while Khmu that resettled recently (N=142) gave an average of 32.0, which is not statistically different from each other (p=0.21 Mann-Whitney U Test). This would corroborate the thesis that their ethnic affiliation provides Khmu in this province with a larger, translocal network. Therefore, the difference between old and new Khmu settlers would be rather insignificant.

As a final step in the analysis, we combine the types of contact participants indicated in their social network with the allocated tokens in the game. Figure 2.6 outlined how often participants mentioned a type of contact in their top 3 across a range of domains. On average, participants allocated 18.4 tokens to these contacts in the game. Table 2.3 compares how often a type of contact was mentioned in the network in all domains with the share of tokens this type of contact received in the game. Family members constituted 43.5% of all contacts mentioned in the survey and received 68.7% of the tokens allocated to the network. Relatives received 22.2% of allocated tokens. Adding these two types of contact together brings the total tokens allocated to the own kin to 90.7%.

| **Type of contact** | **% of times mentioned in the network**<br>(Not counting "Not Filled Out") | **% of tokens allocated to the network**<br>(Only points given to network) |
|---|---|---|
| Family members | 43.5% | 68.7% |
| Relatives | 31.7% | 22.2% |
| Friends | 14.7% | 4.2% |
| Village Chief | 2.7% | 0.6% |
| Ritual Leader | 0.9% | 0.4% |
| Other | 6.5% | 3.9% |

Table 2.3.: Token allocation to network

Given patrilineal succession, unsurprisingly, grown-up sons took priority. One woman in the grandmother's generation explained that she would not give anything to her daughters because they had "gone away" to other houses, while the sons would inherit the house. Also, sons need money and other resources to give as bride wealth, while daughters will be cared for by their (future) husbands. Thus, the patrilineal inheritance system, the obligation to pay bridewealth and the system of virilocal residence that is practised by both Khmu and Rmeet lead to a prioritisation of giving money to sons.

## 2.7. Concluding Remarks

In this article, we have argued for a fruitful combination of social anthropological and experimental economic methods in the study of the distribution of assets. We suggest that

- Behavioural experiments are social, communicative events that can productively be understood in the context of the social relations and communications of the sociocultural field in which they occur;
- Even in such complex ecological situations, experimental games produce data that provide unique insights;
- Certain standard assumptions on which experimental games are based, are less universal than assumed, such as lone decision making.

For the specific case study in Laos, we have demonstrated the embeddedness of the experimental game in social relations such as those with the local administration make the game a unique event, both in comparison to other such games and in terms of the local social landscape. However, it was not only the conditioning – but by no means determination – of the game through local relations of authority and hierarchy that embedded the game in this local context. Many of the procedures to make a game possible, such as recruitment of participants or skills to operate the technology required, depended on the specific circumstances. We also observed the spontaneous emergence of local ways of dealing with the problems that the game posed for people. Thus, the apparently universal way of making decisions on one's own was replaced, in a rather organic way, by more relational ways of solving problems.

Both disciplines aim to reduce the complexity of a given problem or field in their own unique manner. However, their strategies diverge from each other. Economics' reduction of complexity operates through capturing data in the form of numbers and tables, while anthropology focuses on the specificity of life-worlds. While the anthropological approach thus attempts to capture the complexity of a sociocultural situation by focusing on contingent relations and historical specificity, it reduces comparability and generalization. Economics, on the other hand, emphasizes comparability, with an eye to universals, while cutting on the details of life-worlds and local terms, understandings and explanations of behaviour. Thus, while anthropologists emphasize the specificity of contexts, economists tend to focusing on comparison. Both disciplines thus inevitably sacrifice crucial dimensions of social life and scientific precision. A dialogue between them is thus, by necessity, tense and fraught with difficulties to read the respective other, on the one hand, and fertile and enlightening on the other.

The contrasting epistemological approaches of sociocultural anthropology and economics may be summarised in this way:

| | **Sociocultural Anthropology** | **Economics** |
|---|---|---|
| **Methods** | Qualitative > quantitative | Quantitative > qualitative |
| **Complexity reduction** | Reduction to context | Reduction to numbers |
| **Range of datasets** | Small and complex | Large and generalising |

Table 2.4.: Comparison between Sociocultural Anthropology and Economics

# 3. Going Beyond The In-/Out-Group Dichotomy: Investigating Altruism Towards Middle-Groups

## Authors

Leon Houf & Christiane Schwieren [1]

## Abstract

In-group favouritism and out-group hostility are well-known phenomena in research, which are repeatedly documented across fields, contexts and methods. But the world around us does not always fit into this dichotomy of in- and out-group. Often, there is a "middle-group" that shares some identity characteristics with the in-group, but ultimately neither belongs to the in-group nor is an alien out-group. This setting raises the question of how altruistic actions towards a middle-group compare to those in favour of an in-group or against an out-group. Using an online laboratory experiment, we investigate altruistic actions in an allocation task with the minimal group paradigm. Through a between-subject design, we causally find that under two conditions (sufficient sharing of identity marker with the in-group and experience within the experimental task), such a middle-group is treated fairly, creating a clear hierarchy of groups that goes beyond a dichotomy of groups. With this paper, we extend the current research on in- and out-groups conceptually and our results help to model and experimentally test dynamics behind altruistic actions in more complex group settings.

## Keywords

Altruism, Identity, Group Dynamics, Rule Breaking, Multi-Lab & Lab-In-Zoom

## 3.1. Introduction

Humans are social creatures that tend to organise themselves in groups. This can take many forms, from families and tribes to football clubs, shared-flat communities, or fully online communities, for example, on Discord. In very formal organisations and companies, groups or departments are established as well. Looking at this pattern of group formation, central questions in the social sciences are: How do groups regulate themselves? When and how do they remain stable? What are the factors that keep groups stable?

[1] **Status:** This paper was presented at the ESA conference in Bologna in September. Submission to a journal in experimental research is in preparation.
**Acknowlegments:** We thank Ming Dai, Leonie Loy and Patricia von Mellenthin for help in the data collection; Tamas Olah for supporting the oTree implementation; Vinicius Ferraz, Wolf Guardian and Thomas Pitz for enabling the data collection at Rhine-Waal University of Applied Sciences; Melina Hühn and Guido Sprenger for complementing this research project through ethnographic interviews; and Eugen Dimant, Christina Rott and participants at ESA Bologna 2022 for fruitful discussions.

One identified mechanism for stabilising groups is in-group favouritism. In this mechanism, we observe that the in-group is treated favourably, creating long-term benefits for in-group members (Fu et al., 2012). In addition to in-group favouritism, a distinct out-group is often treated unfavourably or even with hostility. This mechanism of out-group hostility is often found to stabilise (in-)group identity yet is not necessarily the inverse of in-group favouritism (Allport, Clark, & Pettigrew, 1954; Brewer, 1999; **?**). When both in-group favouritism, as well as out-group hostility, can be observed, this is identified as parochial altruism (Bernhard, Fischbacher, & Fehr, 2006; Yamagishi & Mifune, 2016). This raises the question of how "parochial" such altruism might be and whether it can be extended to not in-group members (Imada, 2019). In-group favouritism and out-group hostility are well-known phenomena in research, studied in a variety of ways and contexts (Ciccarone, Di Bartolomeo, & Papa, 2020; Goette, Huffman, & Meier, 2012; Li, 2020; Rusch, 2014), often confirmed by the minimal group paradigm (Y. Chen & Li, 2009; Tajfel, 1978) and even in fully remote contexts (Amichai-Hamburger, 2005; Janneck, Bayerl, & Dietel, 2013).

However, not every setting fits into a dichotomy of just an in-group and an out-group. There are many scenarios with a group "in the middle" of a clear and bounded in- and out-group, where the middle-group shares some identity markers with the in-group yet is clearly distinct from it. For instance, we can think of corporate workplaces, where a sales team forms a clear and bounded in-group in contrast to their competitors, the out-group. Here, other departments of the same company (e.g., finance) share characteristics and goals with the in-group, yet these departments are clearly neither part of the in-group of the sales team nor an out-group as a competing company. Similarly, academic institutes or faculties of one particular discipline might regard interdisciplinary teams within their university as a middle-group compared to other disciplines as out-group. Many more examples can be constructed where a simple dichotomy of in- and out-group falls short.

Here, the concept of intersecting identities becomes crucial. Brewer (1999) introduced the idea that individuals can simultaneously belong to multiple groups, each with its own distinct identity markers and social dynamics. This concept, known as "multiple group membership," acknowledges that people often navigate complex social landscapes where their affiliations are not limited to a single in-group or out-group but include various overlapping groups. Such multiple group memberships can influence the intensity and direction of in-group favouritism and out-group hostility. This highlights the need to consider a more layered understanding of social identities to fully grasp group regulation dynamics.

In such group settings, the effects of the regulating forces of in-group favouritism and out-group hostility on the middle-group are a priori unclear, and studies on similarities of groups or with third parties have yielded mixed results (Alves, Koch, & Unkelbach, 2017, 2018; Attanasi, Hopfensitz, Lorini, & Moisan, 2016; Goerg, Hennig-Schmidt, Walkowitz, & Winter, 2016; Kranton, Pease, Sanders, & Huettel, 2020; Linville, Salovey, & Fischer, 1989; Restrepo-Plaza & Fatas, 2022; Yamagishi, Jin, & Kiyonari, 1999). Therefore, compressing all such interactions of groups into an in-/ out-group dichotomy might conceal the underlying behavioural factors that regulate group stability.

In this paper, we present a formalisation of such group settings. Every individual will only belong to one group. This in-group is constructed such that all group members fulfil the same identity markers. The out-group differs in all markers. The middle-group will share some, but not all, of the identity markers. Our central research question is how altruistic actions towards a middle-group compare to in- and out-group. We investigate this question through a multi-lab, lab-in-zoom experiment where subjects are grouped by three types of criteria: arbitrary painting preferences based on the minimal group paradigm, real associations outside the laboratory and random allocation to a group within a session. Subjects perform a 60-round die-rolling task under complete privacy. In this die-rolling task, subjects are given a rule to allocate tokens based on the outcomes of the die-rolling task. They can easily break these rules, as they cannot be observed or punished. We operationalise the social preferences as rule-breaking behaviour to benefit members of specified groups, such as their own experimental in-group.

Our results confirm the general finding of in-group favouritism, where the in-group is always treated favourably. Due to the zero-sum nature of our experimental game, the out-group consistently bears this loss. When a middle-group shares sufficient identity markers with the in-group, and the experimental subjects have already been doing the task repeatedly, then the middle-group is indeed treated neutrally according to a fair allocation rule. Still, the in-group is treated favourably, showing a clear instance of group hierarchy that goes beyond a dichotomy. When a middle-group shares few identity markers with an in-group, it is treated just like an out-group, even when subjects have already been doing the task repeatedly.

Our findings have two main implications for understanding human behaviour across different group settings. First, we document that subjects see beyond a simple in-/out-group dichotomy. When there is a middle-group that shares sufficient identity markers with the in-group, the in-group is treated favourably, the middle-group fairly, and the out-group bears the loss. When a middle-group does not show sufficient identity markers, it is treated like another out-group, i.e., following a clear ‘us-versus-them’ attitude. The purpose of this paper is to test and document that when a group shares sufficient identity markers with an in-group compared to an out-group, it is treated fairly as a middle-group. We thereby show that research should use designs that generally go beyond the in-/out-group dichotomy. The purpose of this paper is not to investigate which identity markers constitute a threshold when an out-group becomes a middle-group, but we can document that in our setting it is future research should investigate this further. Secondly, we find that subjects change their behaviour over the 60 rounds in our game, even when the task and group settings remain constant for the entire duration of the experiment. This shows that findings from experimental designs studying group behaviour should not be overgeneralised as they might only illustrate the short-term effects of an intervention, especially in more complex group settings.

The remainder of the paper is structured as follows: Section 3.2 outlines a framework for identity markers and group settings. Section 3.3 presents the design, and section 3.4 presents the results of our experiment. Section 3.5 concludes.

## 3.2. Framework

The concepts of in-groups and out-groups are well-established in the context of social interactions. However, the real-world dynamics of group interactions often present scenarios that do not neatly fit into these binary categories. This section introduces a model to analyse group settings with multiple identity markers, allowing us to study the nuances of intersectionality in group settings.

### 3.2.1. Identity Markers and Group Definitions

Our model uses identity markers (characteristics) to define groups. For simplicity, we use three markers in the experiment: $x_1$, $x_2$, and $x_3$. Furthermore, we restrict the current model to cases where groups can either share an identity marker $x_i$ or explicitly differ in it, denoted $x'_i$. Using these markers, we can define:

- **In-Group (I)**: Characterized by the identity markers $x_1$, $x_2$, and $x_3$. Formally, $I = [x_1, x_2, x_3]$. This group defines itself as the primary group or the "us" in any given scenario.
- **Out-Group (O)**: Defined as the group that contrasts in all significant identity markers when compared to a member of the in-group. Formally, $O = [x'_1, x'_2, x'_3]$, where each $x'$ represents an identity marker contrasting with the corresponding marker in the in-group.

### 3.2.2. Introduction of Middle-Groups

Between in- and out-group, we introduce the concept of "middle-groups". These groups share some identity markers with the in-group but differ in others and, therefore neither belong to the in-group, nor to the contrasting out-group:

- **Middle-Group A ($M_A$)**: Shares two identity markers with the in-group and contrasts on one. Formally, $M_A = [x_1, x_2, x'_3]$.
- **Middle-Group B ($M_B$)**: Shares one identity marker with the in-group and contrasts on two. Formally, $M_B = [x_1, x'_2, x'_3]$.

### 3.2.3. Group Placement on a Continuum

With these definitions in place, we can visualise these groups on a one-dimensional continuum, as illustrated in table 3.1. This table provides a visual representation of the relative positions of each group based on shared and contrasting identity markers. On this one-dimensional continuum, the in-group is placed at the most left, the Out-Group at the most right position and the remaining groups in the middle. In this continuum, we place the middle-group $M_A$ closer to the in-group, as it shares a greater absolute number of identity markers with the in-group.

| In-Group | | Middle-Group | | Middle-Group | | Out-Group |
|---|---|---|---|---|---|---|
| $I = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ | $\leftrightarrow$ | $M_a = \begin{bmatrix} x_1 \\ x_2 \\ x_3' \end{bmatrix}$ | $\leftrightarrow$ | $M_b = \begin{bmatrix} x_1 \\ x_2' \\ x_3' \end{bmatrix}$ | $\leftrightarrow$ | $O = \begin{bmatrix} x_1' \\ x_2' \\ x_3' \end{bmatrix}$ |

Table 3.1.: Group Placement on Continuum

## 3.3. Experimental Design

In this study, we aim to identify social preferences towards experimental groups. To achieve this, we control the identity characteristic assignment. This allows us to use causal identification methods to answer our research questions. We implement the study as "lab-in-zoom", a multi-lab study programmed in oTree (D. L. Chen, Schonger, & Wickens, 2016).

Our study sample consisted of $N = 244$ subjects from Heidelberg University (52%) and Rhine-Waal University of Applied Sciences (48%). The sample was composed of individuals who were 54% female, 43% male, and 2% non-binary, with an average age of 25.7. Participants represented 41 different nationalities, primarily German (59%) and Indian (9%). The major fields of study were Economics (21%) and Engineering (10%). About 25% were participating in an experiment for the first time, while 19% had participated in more than five experiments. The experimental sessions took place in July and August 2022. On average, subjects earned 10.23€ and spent 31 minutes on the experiment.

### 3.3.1. Group Assignment and Composition

We assign group membership based on several characteristics $x_i$. The first characteristic $x_1$ is created through **arbitrary preferences** on a Klee versus Kandinsky painting, based on the procedure outlined in Y. Chen and Li (2009). The second characteristic $x_2$ is created through the **lived experience** of university affiliation corresponding to the lab where the respective session is conducted. Finally, within a specific session at the same university and having chosen the same painting, groups of three are built randomly. Within such a group, participants jointly solve a group task described below. Characteristic $x_3$ is therefore created through the **randomness** of being in the same session and group and solving a task together.

#### Creation and Definition of In-Group

The in-group is formed following the same procedure within each session. Each session is specific to one university, $x_2$. Subjects initially indicate their preference for a Klee or Kandinsky painting, $x_1$ (see Appendix **??**). They are then randomly put together in groups of 3 based on their shared painting preference.

A 'group task' is then presented, in which subjects still make their decisions individually without any communication with their fellow group members. The purpose of the task is to create a feeling of 'group success'. In this task, subjects are presented with two paintings, one from Klee and one from Kandinsky, and they have to identify which one was painted by the artist they chose previously. This happens over five rounds with new paintings in each round. If the group in total identifies more paintings correctly than incorrectly, the group together has solved the task and all group members receive the same monetary reward. The task is meant to be non-trivial but relatively easy to solve as it only requires a correct identification of the artist slightly better than random guessing. And indeed, in our experiment, every group solves the task. On average, individuals identify 4.5 out of 5 rounds correctly and $\approx 97\%$ of subjects correctly identify at least 3 paintings. While the task is individual, it reinforces group salience, adhering to the minimal group paradigm, where no actual group interaction occurs.

In the context of lab-in-Zoom, group membership remains anonymous as in the standard physical lab situation. The criteria for group formation are clearly defined and thereby bind group membership through the characteristics outlined above.

### Creation and Definition of Out-Group

The out-group is constructed as a group of distant strangers who differ in all identity markers. Participants are told that members of this group have picked the other painting (denoted $x'_1$), are taking part in the experiment at another time, and are members of another university, where no further information is given on that university ($x'_2$, $x'_3$).

### Creation and Definition of Middle-Groups

Furthermore, we create a middle-group that shares some characteristics with the in-group but is clearly not part of the in-group. It is not part of the in-group because the members are not part of the same session and, therefore, do not take part in the group task and also could not have been randomly assigned to it ($x'_3$). The characteristics that are shared are picked painting ($x_1$) and university affiliation ($x_2$). In a variation of the more distant middle-group ($M_B$) we only use the chosen painting as shared characteristics and describe the members of this group as members "of another university".

| | | **In** | **Middle$_A$** | **Middle$_B$** | **Out** |
|---|---|---|---|---|---|
| $x_1$ | **Chosen Artist** | $\surd$ | $\surd$ | $\surd$ | |
| $x_2$ | **University** | $\surd$ | $\surd$ | | |
| $x_3$ | **Session & Group Task** | $\surd$ | | | |

Table 3.2.: Overview group characteristics

Throughout the experiment, we never refer to groups as "in-", "middle-", or "out-group". We always only describe their features (e.g. a member of a middle-group $M_A$ as "someone at *[same university]*, from a different session who chose *[same artist]*".

## 3.3.2. Altruistic Action

The term "altruistic behaviour" describes a wide range of behaviours. In this study, we focus on a specific kind: the act of breaking a small, unenforceable, and unobservable rule for the benefit of another individual. This allows us to investigate altruism under conditions where behaviour is observed rather than self-reported and where the act has no impact on the individual's reputation, removing social incentives. Therefore, there is no personal material gain from the altruistic act, either directly or indirectly, even when directed towards an in-group member.

### 3.3.3. Task

To operationalize our focus on small rule-breaking for the benefit of others, we adapt the "Mind Game" or resource allocation game, previously used to explore moral decision-making and cheating behaviour (Greene & Paxton, 2009; Hruschka et al., 2014; Jiang, 2013; Purzycki et al., 2016). As step 1, participants are instructed to mentally choose a cup as represented in table 3.3. Here, depending on their current round and treatment, they either only see two cups of what we refer to as in- and out-group or three cups including a middle-group. As step 2, they are instructed to roll a die (either physically at home or through an app or website). As step 3, they are instructed to then allocate a coin based on rules that prescribe into which cup the coin should be allocated based on the mentally chosen cup and the outcome of the die role. As example for two cups, if the die shows 1, 2 or 3, they should put the coin into the cup they chose in step 1, if otherwise in the other cup. The rules for 3 cups are outlined in appendix 3.6. Thereby, subjects can always follow or break the rule since nobody can observe the cup they chose or the outcome of the die roll (if they actually rolled a die at all).

This process is repeated 30 times in part 1, and 30 times in part 2, where in some treatments the group composition changes in part 2 as will be explained in the next section.

This design satisfies our criteria for studying small, unenforceable, and unobservable rule-breaking. The procedure and rules make it impossible to identify a single decision as rule-following or rule-breaking. But because of the task's stochastic nature, we can test the aggregate results against the theoretical, rule-following behaviour. Through this comparison and statistical analysis, we can measure treatment effects. The rules were presented very explicitly, and the complexity of the multi-step procedure might increase the cognitive load of participants, which has been found to lessen the propensity to lie (Leib, Köbis, Soraperra, Weisel, & Shalvi, 2021). Furthermore, a large portion of our sample reports a nationality where English is not the official language, which potentially might further increase cognitive deliberation and thereby also lessen the propensity to lie (Bereby-Meyer et al., 2020).

### 3.3.4. Procedure and Treatments

**Session Procedure:**

At the start of the session, subjects are provided a link to the experiment environment on oTree via Zoom, where audio, video and general chat are disabled for participants. In the experiment, participants first make a choice whether they prefer a Klee or a Kandinsky painting. Based on their painting preferences, subjects are randomly grouped and then, as a group, randomly assigned to a treatment condition. Each group carries out their group task. Following this, instructions for part 1 are provided, and subjects individually engage in the task for 30 rounds at their own pace. Subsequently, they receive instructions for part 2, engaging in another set of 30 rounds. Eventually, subjects answer a demographic questionnaire.

**Treatments:**

Our study incorporates three distinct cup setups:

- **In / Out:** This setup involves a cup labelled for the in-group and another labelled for the out-group.
- **In / Middle$_A$ / Out:** This setup includes a third middle cup for group $M_A$, alongside the in-group and out-group cups.
- **In / Middle$_B$ / Out:** Similar to the previous setup, this version includes a cup for the more distant middle-group $M_B$.

Table 3.3 shows these cups as they were presented to the participants. The group concepts in-, middle- and out-group were never mentioned themselves but described as outlined above. The labels #, $*$ and <> were used to identify the individual cups.

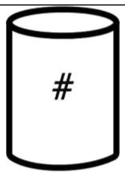
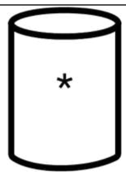
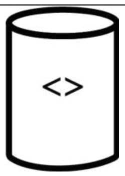
| Group Concept | In | (Middle$_A$) | Out |
|---|---|---|---|
| **Participant View** |  |  |  |
| **Description for Participants** | Someone from your group in this session at UNI where everyone chose ARTIST. | Someone at UNI, from a different session who chose ARTIST. | Someone not at UNI and not in this session and who chose the other painting. |

Table 3.3.: Cup association with group concepts

Table 3.4 provides the overview of our treatment conditions.

- **T1 "2-2":** Is a stable group setting consisting of an in- and out-group.
- **T2 "2-3":** Starts with an in- and out-group in part 1, and additionally introduces a middle-group $M_A$ in part 2.
- **T3 "2-3$_B$":** Also starts with an in- and out-group in part 1, and additionally introduces the more distant middle-group $M_B$ in part 2.
- **T4 "3-3":** Is a stable group setting consisting of an in-, middle- and out-group.

| **Treatment** | **Rounds 1 - 30** | **Rounds 31 - 60** |
|---|---|---|
| **T1** "2-2" | In / Out | In / Out |
| **T2** "2-3" | In / Out | In / Middle$_A$ / Out |
| **T3** "2-3$_B$" | In / Out | In / Middle$_B$ / Out |
| **T4** "3-3" | In / Middle$_A$ / Out | In / Middle$_A$ / Out |

Table 3.4.: Treatment Overview

### 3.3.5. Operationalisation of Rule Breaking

To examine rule-breaking behaviour, we record the number of coins placed into a cup by a participant and compare it to the expected number of coins for each part of 30 rounds. The expected number of coins is 50% of the total coins for two cups and 33% of the total coins for three cups, i.e. 15 and 10 coins over 30 rounds.

As our variable of interest, we measure the extent of deviation from expected values using the percentage deviation, $\%\Delta coins$. A $\%\Delta coins$ value of 0.1, for example, indicates that for every coin that should have been placed in a particular cup, an additional 10% of coins were placed in that cup by that participant.

## 3.4. Results

### In-Group Favouritism Consistent Over All Treatment Groups

First, we measure in-group favouritism as the $\%\Delta coins$ that go additionally to the in-group. Looking at figure 3.1, we see that for all four treatments on the x-axis, for both part 1 and 2, the mean and first standard error of the variable $\%\Delta coins$ is always positive. And besides part 1 in treatment (T1), it is always statistically significant from 0, the theoretical benchmark.[2] Here it is important to note, that treatments (T1) 2-2, (T2) 2-3 and (T3) 2-3$_B$ are identical in this part and we can therefore pool those observations and reach a joint p-value of $p < 0.001$ with a Wilcoxon Signed-Rank Test against the 0 benchmark.

This tells us that irrespective of the composition and timeline of the group setting developed in table 3.1, we always observe in-group favouritism expressed in rule-breaking for the benefit of the in-group. Furthermore, the mean in part 2 exceeds that of part 1 for all treatments except for treatment 4 (the "3-3" treatment), which we will discuss further when we analyse the effect of the middle-group.

### T1: Stable In- and Out-Group Over Time

In treatment 1, the group composition stays stable with an in- and out-group in both parts, as illustrated in table 3.5.

Figure 3.2 shows the averages of $\%\Delta coins$ for all treatments and part for in-, middle-, and out-group. Within each part per treatment, the means of the groups add up to 0, as the

[2] In this paper we use the convention $*$ for $p < 0.05$, $**$ for $p < 0.01$, $***$ for $p < 0.001$.

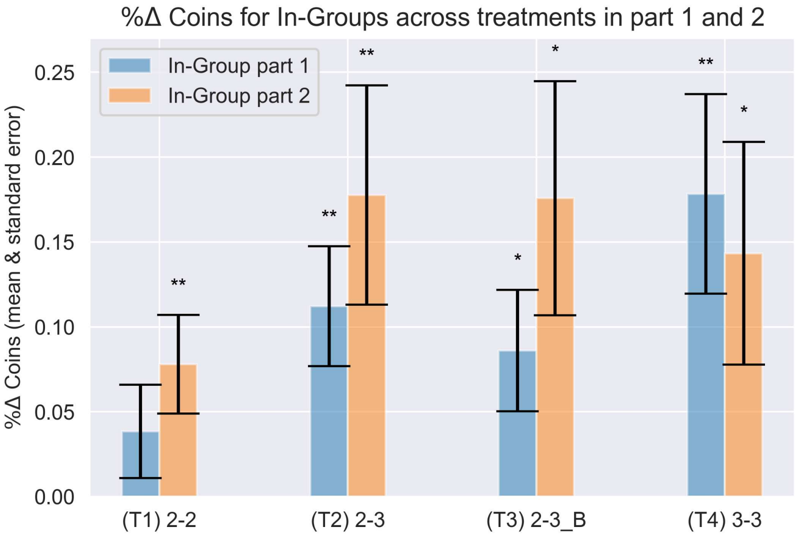


Figure 3.1.: Mean and standard error of $\%\Delta coins$ for in-groups in each treatment over first and second part

| In-Group | | Out-Group |
|---|---|---|
| $I = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ | $\leftrightarrow$ | $O = \begin{bmatrix} x'_1 \\ x'_2 \\ x'_3 \end{bmatrix}$ |

Table 3.5.: Group Setting: In- & Out-Group

| In-Group | | Middle-Group | | Out-Group |
|---|---|---|---|---|
| $I = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ | $\leftrightarrow$ | $M_a = \begin{bmatrix} x_1 \\ x_2 \\ x_3' \end{bmatrix}$ | $\leftrightarrow$ | $O = \begin{bmatrix} x_1' \\ x_2' \\ x_3' \end{bmatrix}$ |

Table 3.6.: Group Setting: In-, Middle- & Out-Group

allocation task is a zero-sum game. For (T1), shown in the top left panel of figure 3.2, $\%\Delta coins$ for the in-group increases from 3.8% to 7.8% in part 2. This indicates a stronger in-group favouritism over time. The value for the out-group is here the inverse, because of the zero-sum nature of the game.

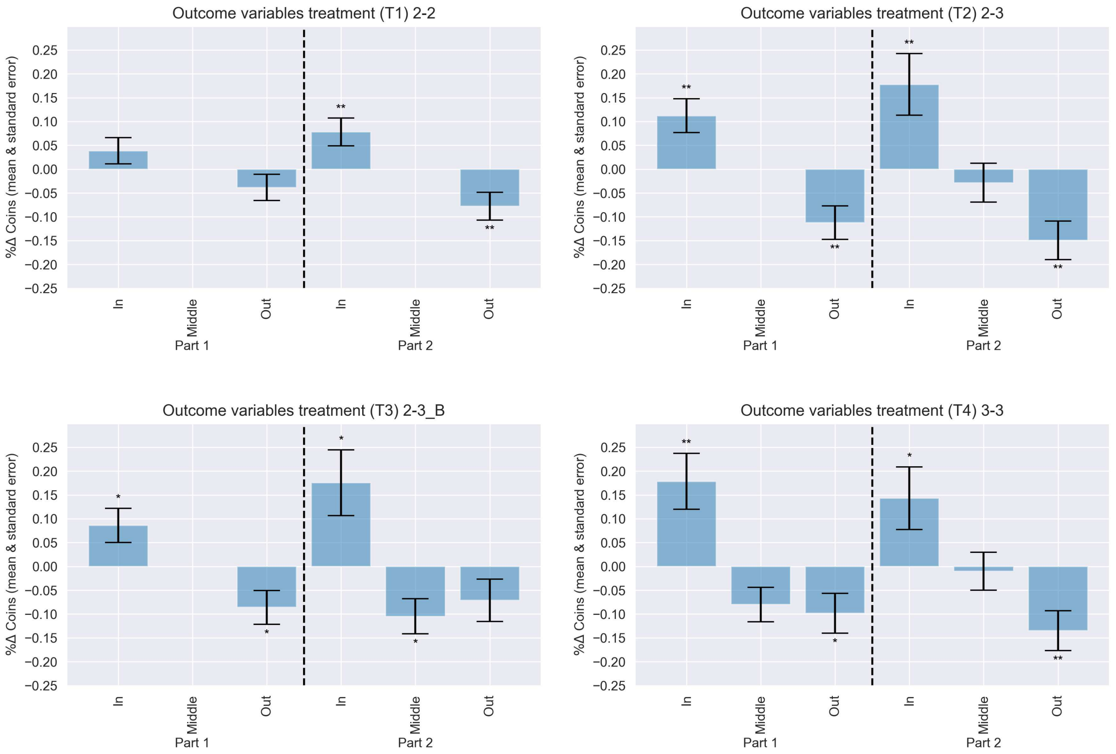


Figure 3.2.: $\%\Delta coins$ for all treatments

### T2: Introduction of a Middle-Group: A Clear Hierarchy of Groups Emerges

In treatment 2, the group composition starts with an in- and out-group in part 1, as in table 3.5. In part 2, a middle group is introduced that shares two identity markers with the in-group, as illustrated in table 3.6. These identity markers are the arbitrary preferences on Klee vs Kandinsky painting and the lived reality of university affiliation.

Figure 3.2 shows in the top right panel for (T2) that $\%\Delta coins$ for the in-group increases from 11.2% to 17.8% in part 2. For the out-group it decreases from -11.2% to -14.9%. In part 2, this is not the inverse because of the presence of the middle-group with a mean of -2.9% in part 2, which is not significantly different from 0.

This shows that when a middle-group that shares sufficient identity markers is introduced,

| In-Group | | Middle-Group | | Out-Group |
|---|---|---|---|---|
| $I = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ | $\leftrightarrow$ | $M_b = \begin{bmatrix} x_1 \\ x_2' \\ x_3' \end{bmatrix}$ | $\leftrightarrow$ | $O = \begin{bmatrix} x_1' \\ x_2' \\ x_3' \end{bmatrix}$ |

Table 3.7.: Group Setting: In-, more distant Middle- & Out-Group

a clear hierarchy of groups emerges, which even strengthens in-group favouritism. As we will discuss below, a question for further research is what a "sufficient" sharing of identity markers is.

### T3 (2-3$_B$): Introduction of a More Distant Middle-Group: Clear "Us (In-Group) versus Them (All Other Groups)"

In treatment 3, the group composition also starts with an in- and out-group in part 1. Then, in part 2, a middle-group that shares only one identity marker is introduced, as illustrated in table 3.7. The shared identity marker is the shared painting preference, while they differ in the other markers, i.e., university affiliation and therefore session.

Similarly as in treatment 2, the bottom left panel in figure 3.2 shows an increase for the in-group from 8.6% to 17.6%. Yet the introduced middle-group receives a $\%\Delta coins$ of -10.5%, which is significantly different from 0, unlike in treatment 2. At the same time, the value for the out-group slightly increases from -8.6% to -7.1%. This indicates that both the more distant middle-group and the out-group are perceived as out-groups. Behaviour towards these groups is not differentiated based on their relative proximity. This can be summarised as 'us (in-group) versus them (all other groups)' behaviour.

This suggests a threshold exists: groups are perceived as more neutral, as in treatment 2, when they share enough identity markers. Conversely, in cases like treatment 3, 'insufficient' shared identity markers result in all non-in-groups being perceived as out-groups. The purpose of this paper is not to develop and test a threshold when exactly a group is perceived as middle- and when as out-group, but to document that it exists. The difference between (T2) and (T3) is the presence of the identity marker of university affiliation, i.e. lived reality. From this setting, further research should start by investigating how identity markers of lived reality can change perception and thereby associated altruistic behaviour of an out-group to a middle-group.

### T4: Stable, In-, Middle- and Out-Group

In treatment 4, the group composition stays stable over time, including a 'closer' middle-group that shares two identity markers of painting preferences and university affiliation, as illustrated in table 3.6.

The bottom right panel in figure 3.2 shows that in part 1, the in-group receives 17.8% in $\%\Delta coins$, while the middle- and out-group receive -8.0% and -9.8%. This again can be described as an 'us (in-group) versus them (all other groups)' behaviour. This suggests that in the beginning of a new environment, experimental subjects do not differentiate more than a simple in- and out-group effect. Yet in part 2, they clearly do. Here, the in-group receives 14.3%, the middle-group receives -1.0%, not significantly different from 0, and the out-group receives -13.5%. This mirrors the results of treatment 2 in part 2 and shows a clear hierarchy in terms of relative proximity for groups after experimental subjects have become used to the task.

## 3.5. Discussion & Conclusion

In this study, we have investigated altruistic rule-breaking across various group settings. We designed and controlled these group settings in a multi-lab, lab-in-zoom study, using identity markers derived from real-world associations, the minimal group paradigm and randomness within the lab. This allowed us to create an in-group, an anonymous out-group of distant strangers and middle-groups who share some identity markers with the in-group, yet are clearly distinct from it. The altruistic rule-breaking in a die roll game in complete privacy is measured through aggregate analysis.

Our findings confirm the general presence of in-group favouritism. Furthermore, we can report a clear pro in-group behaviour when a middle- and out-group are present in the *initial part* of the experiment. Yet this "us-versus-them" behaviour at the start evolves into a behaviour reflecting a hierarchy of groups in the second part of the experiment (T4). Here, we observe a pro in-group and neutral towards middle-group behaviour compared to a rule-following statistical benchmark. Accordingly, the out-group carries the loss in our zero-sum-game setting.

This fair and neutral behaviour towards a middle-group is also confirmed in Treatment 2, where such a group is only introduced in the second part of the study. At the same time, participants show sensitive behaviour to the composition of identity markers of such a middle-group. When a middle-group is introduced that shares fewer identity markers with the in-group and is lacking the identity marker based on lived experience (university affiliation), then it is treated like an out-group, creating a clear "us-versus-them" behaviour again (T3).

These findings have multiple implications for future experimental research on inter-group behaviour. This study suggests that in one-shot interactions or short experiments, participants would show strong in-group favouritism and less nuanced behaviour towards more complex group settings, yet participants might show more nuanced behaviour later in that exact same study after being exposed to this setting for a longer time. Furthermore, we present a group setting going beyond the classical in- and out-group dichotomy using several identity markers. Employing multiple identity markers to go beyond that dichotomy could significantly benefit research on inter-group relations.

Here, we see a vast potential for further research. In the present study, we see a middle-group, $M_A$, that is treated neutrally, and a middle-group, $M_B$, that is treated just like an out-group. This raises questions on the composition of those identity markers, $x$. How important is each single, isolated identity marker? Which role do markers from lived experience play? How are their conjoint effects? Is there a tipping point in markers, where a 'neutral' behaviour turns 'hostile', or would it be a gradual transition? Answering such questions builds directly on the framework of the present study and can inform intergroup research and important current social issues around discrimination and polarisation, especially in contexts of intersectionality.

In this study, we transition from a group dichotomy to a linear ordering of groups based on identity markers. Further research should increase this complexity to settings where a marker $x$ can take more than two values and interact with other markers. Additionally, in the present study, the markers of in- and out-group remained constant. Future research should analyse the dynamics of multiple groups and changing markers $x$ within and across groups, potentially with (partial) uncertainty of markers.

Naturally, this present study carries limitations. In the task, we implement a comparably high mental cost of rule-breaking by providing a very strict set of rules. How the results change, especially towards a middle-group, when the mental costs of rule-breaking change (e.g., even lower) is not clear. We can also not know whether experimental demand effects might have directed behaviour in a particular direction. In our study, participants take part in the privacy of their homes. Would it make a difference to run the study in a lab or at a public place? The number of coins to be allocated is fixed and a zero-sum game in our setting, making the results within each experimental part and treatment condition complements to each other, especially in settings of two groups. Lastly, since we are unable to *know* for which exact coin allocation a subject was breaking the rule, we must rely on an aggregate analysis level. While it is possible to detect outliers whose behaviour is very unlikely to have been rule-following, any individual behaviour closer to a rule-following behaviour can never confidently be identified as rule-following or breaking. On the one hand, this true privacy is a huge benefit of the method to elicit true behaviour, but it brings limitations to the analysis. What we know from voluntary, ethnographic interviews conducted after the experiment by Melina Hühn and supervised by Guido Sprenger is that some participants stated that they were fully rule following when the die rolls prescribed to give to the in-group, but when it should go to the out-group multiple times in a row, they would stop that "lucky streak" at some point by giving to the in-group instead.

In conclusion, our study robustly demonstrates that subjects in an experimental setting perceive beyond the simple dichotomy of in- and out-groups. They exhibit more nuanced behaviours towards complex group settings. With our introduction of a middle-group it can lead to a clear hierarchy of altruistic rule-breaking where a middle-group is treated neutrally. This finding challenges the traditional in-/out-group dichotomy and suggests more rich group settings and dynamics. Further research should follow up on this group setting paradigm to understand better situations we observe in the world around us to eventually inform policies tackling polarisation and discrimination.

# Appendix

## 3.6. Rules For Coin Allocation with Three Cups

When participants have to allocate their coin between three cups, step 1 is to choose one of those cups in their mind. Step 2 is to roll a die. In step 3, these are the rules given:

- If the die shows 1 or 2, you will put one ECU in the cup you chose in your mind in Step 1.
- If the die shows 3 or 4, you will put one ECU in the cup that is clockwise to the cup you chose in Step 1 following the BLUE arrows.
- If the die shows 5 or 6, you will put one ECU in the cup that is counter-clockwise to the cup you chose in Step 1 following the RED arrows.

Referring to this image:

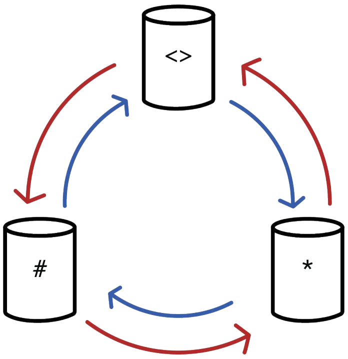


Figure 3.3.: Allocation rule for three cups

# 4. Measuring In-Group Favouritism on Prolific: Experimental Evidence on When Prolific Participants Show Social Preferences

## Author

Leon Houf [1]

## Abstract

This study presents two jointly pre-registered tasks that both aim to measure in-group favouritism. The first task is a novel adaptation of the slider task, which allows to separate in-group favouritism from out-group hostility and to cluster individual behaviour to analyse heterogeneous effects. Essentially, in the present study this task yields a null-result. The second task, a generosity task, clearly shows in-group favouritism. The participants do both tasks after each other and are recruited via Prolific. The fact that one task does not result in in-group favouritism, but the other task does, creates an interesting opportunity to analyse *why* and *when* a task prompts in-group favouritism in the behaviour of the subjects. Based on the tasks, this paper presents as hypothesis that behaviour of Prolific participants follows a hierarchy: **(1.)** They maximise their own bonus payment. **(2.)** They do engage conscientiously with a task, even without extra financial incentive. **(3.)** Only when directly asked to do so, they show social preferences. This hypothesis should be evaluated through future research to increase our understanding of the Prolific subject pool.

## Keywords

Social Preferences, Prolific Participants, Identity

## 4.1. Introduction & Motivation

Prolific is a central subject pool widely used to produce knowledge in behavioural economics and the social sciences in general (Galizzi & Navarro-Martinez, 2019; Palan & Schitter, 2018; Stanton, Carpenter, Nance, Sturgeon, & Andino, 2022). Recent studies with subject pools on Prolific have investigated various aspects of human behaviour, including social preferences, such us dishonesty in online experiments (Parra, 2024) or trust games (Safra, Lettinga, Jacquet, & Chevallier, 2022).

To prevent publication bias and advance our understanding of the data generated by participants on Prolific, it is essential to investigate whether fundamental, underlying

[1] **Status:** The idea to this paper was presented at ASFEE in Lyon in May 2022. Submission to a journal focusing on methods in experimental research is in preparation.

behavioural patterns can be detected in this subject pool. The purpose of this paper is to present a hierarchy of such behaviour that can be inferred through a pre-registered study on in-group favouritism because it only partially yields a null-result.

In this study, two tasks are done by the same participants after each other: a novel adaptation of a slider task stylised as performance review task and a generosity task labelled as giving tip.

The performance review task allows to separate in-group favouritism from out-group hostility. The motivation to do so is that studies have shown that in-group favouritism and out-group discrimination are distinct behaviours driven by different motives. In artificial groups, favouritism is observed without discrimination, suggesting these behaviours are not the same (Abbink & Harris, 2019; Rusch, 2014). In fact, these two behaviours towards groups can take different forms in many work situations such as (contract-) negotiations, bonus payments or performance evaluations (Franco & Maass, 1996; Li, 2020; Smith, DiTomaso, Farris, & Cordero, 2001). Additionally, the nature of in-group favouritism and its enforcement as a social norm can vary depending on the context and the group involved (Y. Chen & Li, 2009; Ciccarone et al., 2020; Dimant, 2024; Harris, Herrmann, Kontoleon, & Newton, 2015). Furthermore, the task is able to cluster and analyse heterogeneity in behaviour, inspired by Kranton et al. (2020) who classify "groupy" and "not-groupy behaviour" in social preferences.

In the performance review task, there are "workers" and "supervisors." Supervisors get either a fixed bonus payment or a bonus based on their reviews which includes risk. Workers perform under a piece-rate scheme or compete against another worker in a winner-takes-all tournament. Each worker experiences both schemes in different parts of the study. To induce social preferences, we recruit participants on Prolific in the U.S. based on their identification as Democrat or Republican. This serves as the operationalisation of in- and out-group based on shared or different party affiliation between supervisor and worker, as used by, for example, Balliet, Tybur, Wu, Antonellis, and Lange (2016) and Dimant (2024).

The results show that within the performance review task, supervisors treat workers from their own and the other party equally. They do not change their behaviour based on the workers' incentive scheme (piece-rate versus winner-takes-it-all). However, supervisors with fixed bonus payments are more lenient to both workers in their reviews than those with risky bonus payments, who become stricter over time. But while they are more lenient, they still conscientiously and consistently engage with the task and treat both workers equally and do not show social preferences. We also identify clusters of participant behaviour that largely remain stable over time. In summary, our findings on social preferences within the performance review task can be described as null-results regarding in-group favouritism.

However, when supervisors have the opportunity to give an additional tip after the performance review, they show clear in-group favouritism by giving significantly and substantially more to workers from their own party. This combination of results provides methodological insights into how participants on Prolific respond to studies that elicit social preferences. Based on this study, we can formulate a hierarchy of behaviour. **(1.)** When their own bonus payment is at risk, they behave according to a risk-neutral manner, showing neither in-group favouritism nor out-group hostility. **(2.)** When their own bonus payment is not affected, they still conscientiously complete the task without demonstrating social preferences. **(3.)** Only when the task explicitly prompts them to show social preferences (such as the opportunity to give a tip) and the task has no impact on their bonus payment, then they exhibit clear social preferences.

The remainder of this paper is structured as follows: Section 4.2 describes the performance review task, its operational validity, and the overall experimental design, section 4.3 presents the results and section 4.4 further discusses the methodological implications for research on social preferences and concludes.

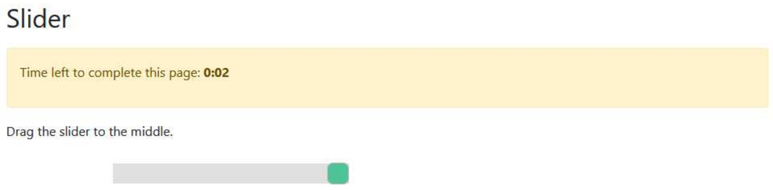
## 4.2. Experimental Design

### 4.2.1. Performance Review Task

The experiment consists of two distinct phases. In the first phase, a group of Prolific participants is assigned the role of "workers" to produce a "product." These results are then evaluated in the second phase by a new and separate group of Prolific participants acting as "supervisors" in a performance review task. Workers are recruited from Prolific, required to be based in the US, and evenly split between self-identified Democrats and Republicans, which is never mentioned and irrelevant to the workers.

**Phase 1: Workers Produce Products (Modified Slider Task)**

In this task, workers engage in a modified slider activity. They are shown a range with a slider initially positioned at the far left or right edge of the range. The task is to centre the slider in that range within a two-second time limit, creating a time-pressured environment. The participants' screen is shown in figure 4.1.



Slider

Time left to complete this page: **0:02**

Drag the slider to the middle.

Figure 4.1.: Workers' screen slider task

This task is repeated 50 times in part 1 and 50 times in part 2. The range's width and position on the page change randomly each round, which is the key challenge for the workers in the task, as it makes it more difficult to determine the exact middle of the range. This variability and the time pressure lead to heterogeneity in outcomes across rounds and workers. These final placement of all sliders that have been moved by the workers can be seen in figure 4.2. It clearly shows a normal distribution around the middle of the range (value 100).

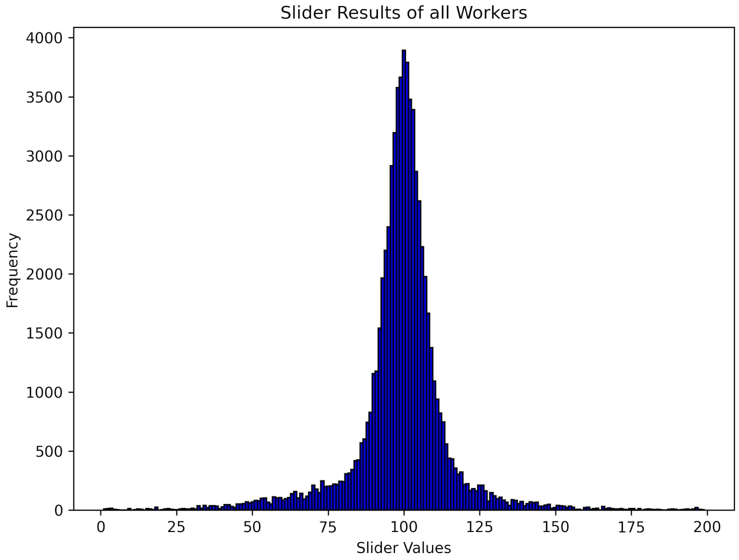
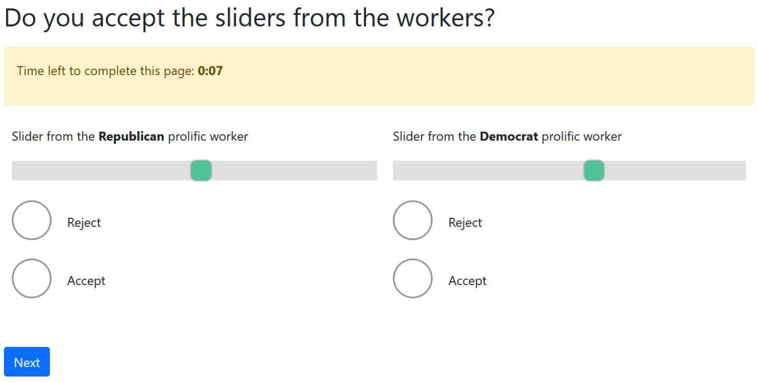


Figure 4.2.: Slider Results all Workers


Do you accept the sliders from the workers?

Time left to complete this page: **0:07**

Slider from the **Republican** prolific worker

○ Reject

○ Accept

Slider from the **Democrat** prolific worker

○ Reject

○ Accept

Figure 4.3.: Supervisors' Screen Decision Task

### Phase 2: Supervisors Evaluate the Workers' Output

The data carried from phase 1 to phase 2 of the experiment are the slider value of each round, assigned treatment, and whether a worker identified as Democrat or Republican. The workers' behaviour in phase 1 is not further interesting for this study. In phase 2, a new set of Prolific participants is recruited, again evenly split between self-identified Democrats and Republicans. All of these participants are assigned the role of a supervisor. Each supervisor is presented with the results of the same two workers throughout the entire experiment. In each round, the supervisor sees the results of where the workers put their slider in the respective round and can decide for each worker whether to “accept” or “reject” this slider as a product. The screen of the supervisors is shown in figure 4.3.

## 4.2.2. Generosity Task of Giving Tip

As pre-registered, at the very end, after all rounds are finished, supervisors have the opportunity to give up to 20 points to the workers, labelled as a “tip”. This does not affect the supervisors' payments. Before deciding how many points to tip, supervisors see the results of the 100 rounds. This includes the worker's party identification (as before), the number of sliders each worker accepted, and the number of complaints about these sliders for each worker.

### 4.2.3. Treatment Structure

The design allows us to measure three levels of manipulation effect: the identity marker of Republican versus Democrat as the operationalisation of in- versus out-group, the impact of supervisors' payment structure, and the effect of workers' payment structure on the behaviour of the supervisors.

#### 4.2.3.1. Identity Marker

All participants are recruited from Prolific based on their self-identification as either Republican or Democrat. To the workers, this aspect was never mentioned to prevent any behaviour from reacting to anticipated discrimination based on their identification.

To the supervisors, the workers' self-identification was used to distinguish the two workers, as to them, always one Republican and one Democrat was shown. This group identification mechanism follows studies such as Balliet et al. (2016) and Dimant (2024). The supervisors' party-self-identification was not mentioned or made salient to prevent experimenter demand effects.

We will test whether supervisors show in-group favouritism and/or out-group hostility based on the workers' party affiliation.

#### 4.2.3.2. Supervisor Payment

Through two treatment settings, we test different incentive structures for the supervisors' additional bonus payments: risky and fixed payment. This bonus payment is also paid via Prolific in addition to the necessary Prolific 'show up fee'.

**Risky Payment**

In the first treatment, the bonus payment the supervisors will receive is determined by a process involving risk. Every slider ("product") a supervisor accepts is sent to an automatic "customer" in this setting. This customer can either accept or complain about a slider. The complaint probability $p$ for each slider is automatically determined by the distance of that slider to the middle of the range, $p = \frac{\text{distance to the middle of range}}{\text{max distance to the middle of range}}$. Thus, a slider perfectly placed in the middle will have a complaint probability of $p = 0\%$ and a slider that was not moved at all from the edges of the range will always yield a complaint probability of $p = 100\%$.

If a supervisor accepts a slider and the customer does not complain, the supervisor receives 1 point. If the customer does complain, the supervisor loses 10 points. Accordingly, the instructions for supervisors clearly state that if they accept a slider with a distance of $\approx 9\%$ or less to the middle, they are more likely to earn money than to lose money. The analysis will consider this 9% threshold as the benchmark for a risk-neutral supervisor with risky bonus payments. If the supervisor rejects a slider, this slider is not subject to this probability calculation.

Supervisors do not get feedback on whether an individual slider was accepted or rejected by the automatic customer. Only at the end, after all 100 rounds, are all points from accepted sliders added, and all points from complaints deducted. If these total points are > 0, each point is paid as a 0.01$ bonus payment via prolific. In our study, this was 0.51$ on average across all supervisors in this risky bonus payment treatment setting.

**Fixed Payment**

In the treatment setting with fixed bonus payments for the supervisors, they are asked to carry out the slider review task just as the other supervisors, but for them, no risky bonus payment is involved. They all receive a 0.51$ bonus payment, just as the participants in the risky payment treatment received on average. This means that whether a supervisor accepts all or none sliders has no payment relevance for the supervisor, but only for the workers. For the workers it would be beneficial if a supervisors accepts all sliders.

#### 4.2.3.3. Worker Payment

Next to the prolific base pay (show up fee), the bonus payment of the participants that were assigned the role of workers has two schemes: piece rate and tournament winner-takes-it-all payment. All workers are subject to both incentive schemes. The first 50 rounds are under piece rate, and the second 50 rounds are used as a tournament, or vice versa. Workers are told that they are paired with another worker and their assigned supervisor continually evaluates the same two workers simultaneously.

**Piece Rate**

Under the piece rate incentive scheme, every worker receives a 0.01$ bonus payment for every slider the assigned supervisor accepts. This is accumulated over the 50 rounds.

**Tournament**

In the tournament scheme, workers are told that they will compete against another worker, and the worker that has more sliders approved by the assigned supervisor gets 0.01$ for every accepted slider of both workers in this winner-takes-it-all tournament.

#### 4.2.3.4. Treatment Overview

Supervisors either have a *risky* or *fix* payment. Workers either have first the tournament incentive scheme and then the piece rate *TP* or vice versa *PT*. The combination yields four treatment conditions summarised in table 4.1.

| | | Supervisor Payment | |
|---|---|---|---|
| | | Risky | Fixed |
| Worker Payment | Tournament - Piece Rate | *Risky-TP (T10)* | *Fix-TP (T20)* |
| | Piece Rate - Tournament | *Risky-PT (T11)* | *Fix-PT (T21)* |

Table 4.1.: Treatment Overview

### 4.2.4. Operational Validity of Method

As explained above, supervisors should reward workers when the slider is close to the middle of the range. Figure 4.4 shows that supervisors are more likely to accept sliders when they are close to the middle value (100). At the margins, they are less likely to accept the sliders, with increased confidence intervals due to fewer sliders having these values (as shown in figure 4.2).

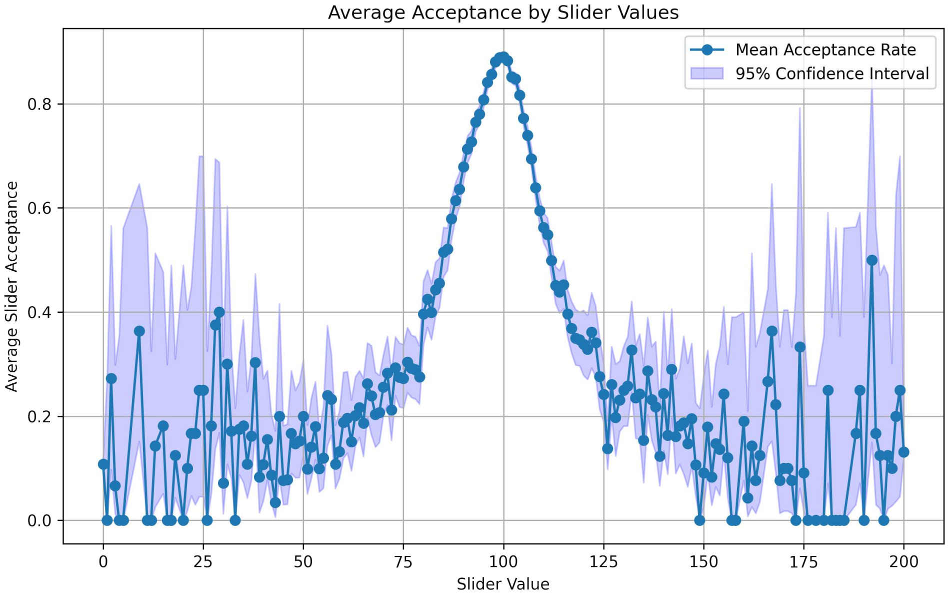


Figure 4.4.: Average Slider Acceptance by Slider Value

As pre-registered (#113497), our main variable of interest is the threshold at which a supervisor is more likely to accept than reject a slider of a given player $i$ in part $t$. We calculate this *Slider-Acceptance-Threshold (sat)* exactly as pre-registered and explained in detail in appendix 4.6. The *sat* represents the distance to the middle of the range on a scale from $0 - 100$ from which on a supervisor was more likely to accept than to reject a slider. To allow for noisy behaviour, it is constructed such that the number of rejected sliders closer to the middle than the *sat* equals the number of sliders accepted further away from the middle than the *sat*. Therefore, the *sat* does not depend on the average performance of a worker, but on the threshold on which a supervisor was more likely to accept a slider. For example, if a supervisor would accept every slider with a distance smaller than 10, and reject every slider with a distance higher than 10, the *sat* score will be 10 for a given worker as long as there is at least 1 slider on either side of 10. The average performance of a given worker does not play a role for the calculation of the *sat*.

The closer a *sat* value is to 0, the more "demanding" a supervisor is, i.e. a supervisor accepts only sliders that are very close to the middle of the range. A *sat* of 0 indicates that only sliders perfectly placed in the middle were accepted. The higher the *sat* score is, the more "lenient" a supervisor is, i.e. accepting sliders that are also further away from the middle. A *sat* close to 100 indicates that almost every slider was accepted. Figure

4.5 shows the *sat* values for all supervisors. The 25th percentile is at a *sat* score of 5, the 50th at 8 and the 75th at 13. The mode is a *sat* score of 5. 62.39% of all *sat* scores have a value of 9 or smaller, which is the threshold for a risk-neutral agent in the risky bonus payment treatments.

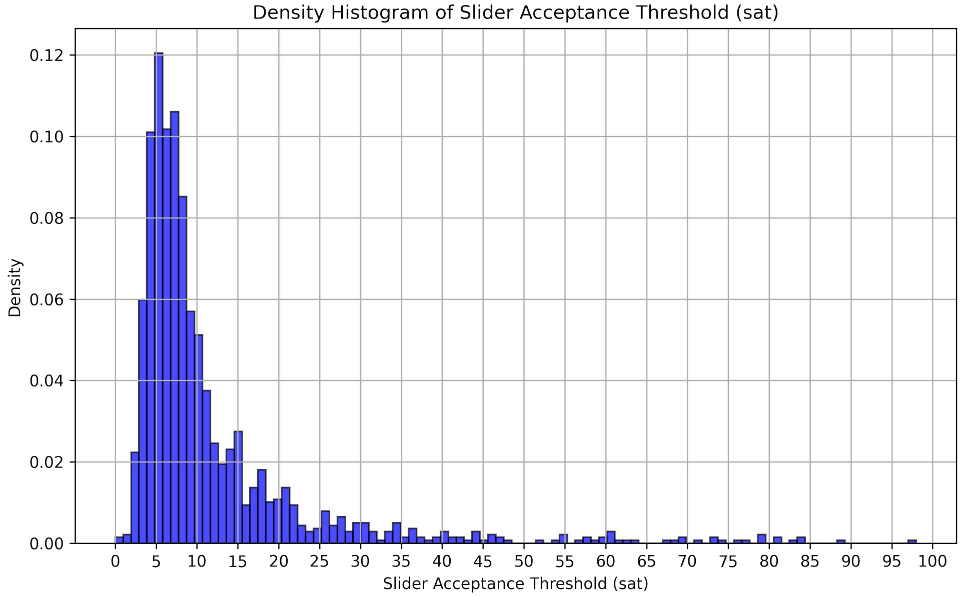


Figure 4.5.: Density Histogram of Slider Acceptance Threshold (sat)

The results of the workers shown in figure 4.2, and for the supervisors in figures 4.4 and 4.5 show that the method fulfils its desired purpose. The task creates variation in behaviour in both roles in well-behaved distributions in line with the incentives of the task.

**Differentiating In-Group Favouritism from Out-Group Hostility**

It is crucial to note that every supervisor will have a *sat* for each worker and that these measurements are independent from each other and do not depend on average performance. A supervisor can be very lenient towards both, one or none of the workers. Therefore, the scores for the workers are not a zero-sum game but can be used to differentiate the supervisors' behaviour towards the two workers. As a benchmark analysis, we can compare those *sat*-scores against the benchmark of a risk-neutral supervisor. Furthermore, we can analyse how the two workers are treated differently from each other.

### 4.2.5. Demographics of Sample

Table 4.2 summarises the categorical demographics of the supervisor sample. There are $N = 357$ supervisors with an average age of 42.6 years. Table 4.3 presents the number of supervisors per treatment. The sessions with risky bonus payment were conducted in November 2022 and the sessions with fix bonus payment in November 2023 and the experiment was programmed in oTree (D. L. Chen et al., 2016).

| **Language** | | **Ethnicity simplified** | | **Country of birth** | | **Employment status** | |
|---|---|---|---|---|---|---|---|
| Count | Value | Count | Value | Count | Value | Count | Value |
| 343 | English | 277 | White | 332 | U.S. | 137 | Full-Time |
| 2 | Spanish | 31 | Black | 3 | Ghana | 57 | Part-Time |
| 2 | Chinese | 21 | Mixed | 2 | India | 38 | Not in paid work |
| 8 | Other | 12 | Asian | 19 | Other | 30 | Job seeking |
| | | 13 | Other | | | 16 | Other |
| **Sex** | | **Student status** | | **Country of residence** | | **Nationality** | |
| Count | Value | Count | Value | Count | Value | Count | Value |
| 178 | Female | 271 | No | 355 | U.S. | 355 | U.S. |
| 177 | Male | 34 | Yes | | | | |

Table 4.2.: Descriptive Statistics for Categorical Variables

| **Treatment** | **Risky-TP (T10)** | **Risky-PT (T11)** | **Fix-TP (T20)** | **Fix-PT (T21)** |
|---|---|---|---|---|
| **N** | 95 | 92 | 82 | 88 |

Table 4.3.: N per Treatment

## 4.3. Results

This section presents the performance review task results from 1 - 3. An additional (pre-registered) analysis on heterogeneous behaviour is in the appendix as result (I). Result 4 analyses the tip task.

### 4.3.1. Result 1: Supervisors do not treat workers significantly differently based on party affiliation

As the first level of analysis, we want to evaluate whether supervisors treated the two workers differently based on their party affiliation. We compare the slider acceptance threshold (*sat*) of a given supervisor for the worker of their own party compared to the other worker as a delta: $\Delta sat = sat_{OwnParty} - sat_{OtherParty}$, a positive $\Delta sat$ represents a higher "leniency", i.e., favouring the worker of the own party.

Figure 4.6 shows the mean $\Delta sat$ and 95% confidence interval for each treatment and part. The mean value is positive (i.e., more leniency for the in-group worker) for all treatments except treatment *Risky-TP (T10)*, which has a slightly negative effect (in part 1 at $-0.01$ and therefore almost not visible in the chart).

Furthermore, we observe that all confidence intervals overlap 0. This shows that at the treatment level, there is no statistically significant in-versus-out-group effect in accepting sliders based on party affiliation.

### 4.3.2. Result 2: Supervisors do not change their behaviour based on the incentive schemes assigned to the workers

We investigate whether the incentive scheme assigned to the workers in part 1 (piece rate versus winner-takes-it-all) influences the *sat* on the treatment level.

First, we compare the treatments with risky incentive schemes for the supervisors. For this, we compare the data of the first part in treatments *Risky-TP (T10)* and *Risky-PT (T11)*. Conducting a two-sided Mann-Whitney U test, we do not find a statistical

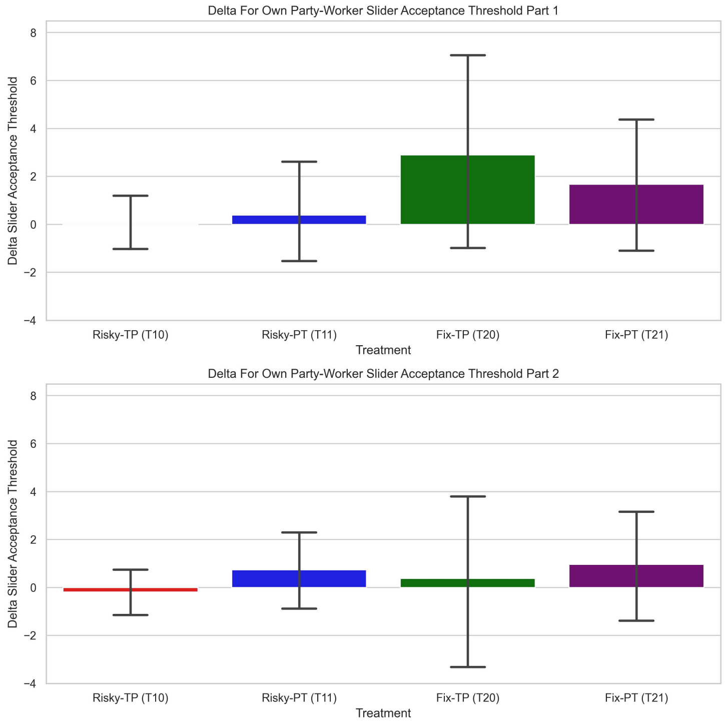


Figure 4.6.: Delta For Own Party-Worker Slider Acceptance Threshold

difference between all *sat* scores in part 1 between treatments *Risky-TP (T10)* and *Risky-PT (T11)* ($p = 0.13$). Conducting the same test on only the *sat* scores for the own party worker ($p = 0.29$) and only on the other party worker ($p = 0.29$) also shows no statistical difference.

The supervisors subject to a fixed bonus payment incentive scheme also do not treat workers differently based on the workers' incentive scheme, which we test between subjects with treatments *Fix-TP (T20)* and *Fix-PT (T21)* and a Mann-Whitney U Test ($p = 0.81$). This also holds for only own party workers ($p = 0.99$) as well as only other party workers ($p = 0.78$).

### 4.3.3. Result 3: Supervisors with risky bonus payments become stricter in part 2

The workers are under different incentive schemes in part 1 and part 2, which is known to the supervisors. Table 4.4 shows for all treatments the *sat* scores for the in-group workers of part 1 in column 1, for part 2 in column 2 and in column 3 this difference as $\Delta = part2 - part1$. Column 4 reports the p-values of a Mann-Whitney U Test between the *sat* scores of part 1 and part 2. Here we can observe that in both treatments with risky supervisor payment, the supervisors became more strict in part 2. Supervisors in treatment *Risky-PT (T11)* start with a *sat* of 11.07 in part 1 and become stricter to almost exactly the risk-neutral threshold at $sat = 9.02$. The supervisors in treatment *Risky-TP (T10)* are, on average, below the risk-neutral threshold. The two treatments are not significantly different from each other in their $\Delta$ values; this suggests that supervisors become stricter not because of the incentive schemes the workers are subject to but because they become more familiar with the task and become more strict over time because of their own incentive scheme. This is supported by the fact that the supervisors with fixed bonus payment,*Fix (T2x)*, have no significant change in their *sat*, and they are always well above the risk-neutral threshold of a *sat* of 9.

| | | *sat* scores | | | |
|---|---|---|---|---|---|
| | | part 1 | part 2 | $\Delta$ = p2 - p1 | p-value |
| Treatment | Risky-TP (T10) | 8.76 | 7.51 | -1.26 | $0.006^{**}$ |
| | Risky-PT (T11) | 11.07 | 9.02 | -2.05 | $0.001^{**}$ |
| | Fix-TP (T20) | 14.90 | 15.06 | 0.16 | 0.433 |
| | Fix-PT (T21) | 14.02 | 13.98 | -0.04 | 0.497 |

Table 4.4.: *sat* scores across parts

### 4.3.4. Result 4: Supervisors give significantly and substantially more tip to workers from the own party

At the end of the 100 rounds of reviewing sliders, supervisors can give up to 20 points labelled as ‘tip’ to the players. Supervisors can freely allocate this budget. For example, they could decide to give no tip at all or give 10 points to each, 20 to one worker and nothing to the other or any other combination that adds up to a maximum of 20 points. Each point is converted to 0.01$. The workers had no information on this and just received it is an additional unexpected bonus payment. Giving (or not giving) tip has no payment effect on the supervisors themselves. On average, supervisors tip a total of 17.8 points to both players. 86.83% of supervisors use the entire budget of 20 points, and 7.56% of supervisors do not give tip at all. Before giving the tip, supervisors see the aggregated results of the 100 rounds: for each player, how many sliders they accepted, how many complaints occurred on these sliders and the party-identification of the players exactly as during the task.

We want to analyse whether supervisors show in-group favouritism towards the workers from their own party by giving them more tip. Table 4.5 presents the results of an OLS regression with *tip per worker* as the dependent variable and *own party worker* as the main variable of interest. We also include control variables such as the treatment, sex, and political party of the supervisor, whether a worker wins the tournament, the number of accepted sliders, and the number of complaints for each worker. The standard errors are clustered at the supervisor level.

We observe the intercept at 7.56 points per slider. Our main variable of interest, *own party worker*, shows that supervisors give 3.17 points more to the workers of their own party. This result is both statistically significant ($p < 0.001$) and also economically significant given the large effect size and shows clear in-group favouritism towards the worker of the own party. Furthermore, we find that supervisors in the fixed bonus payment treatments give significantly more tip ($\approx 1.2$ points). We also find that the more sliders were accepted, the more tip was given to a worker. This is equally counterbalanced by the negative coefficient for accepted sliders of the other worker.

| | | | |
|---|---|---|---|
| **Dep. Variable:** | tip per worker | **R-squared:** | 0.194 |
| **Model:** | OLS | **Adj. R-squared:** | 0.183 |
| **Method:** | Least Squares | **F-statistic:** | 14.54 |
| **No. Observations:** | 794 | **Covariance Type:** | cluster (per supervisor) |

| | **coef** | **std err** | **z** | **P> \|z\|** | **[0.025** | **0.975]** |
|---|---|---|---|---|---|---|
| **intercept** | 7.5596 | 0.736 | 10.273 | 0.000*** | 6.117 | 9.002 |
| **own party worker** | 3.1746 | 0.417 | 7.611 | 0.000*** | 2.357 | 3.992 |
| **Treatment: Risky-PT (T11)** | -0.1737 | 0.496 | -0.350 | 0.726 | -1.147 | 0.799 |
| **Treatment: Fix-TP (T20)** | 1.2662 | 0.381 | 3.320 | 0.001** | 0.519 | 2.014 |
| **Treatment: Fix-PT (T21)** | 1.1815 | 0.399 | 2.961 | 0.003** | 0.399 | 1.964 |
| **Sex: Male** | -0.3889 | 0.287 | -1.357 | 0.175 | -0.951 | 0.173 |
| **Politics: Democrat** | 0.0840 | 0.286 | 0.293 | 0.769 | -0.477 | 0.645 |
| **worker wins tournament** | -1.0439 | 0.524 | -1.991 | 0.046* | -2.071 | -0.016 |
| **accepted sliders** | 0.0753 | 0.012 | 6.335 | 0.000*** | 0.052 | 0.099 |
| **other accepted sliders other worker** | -0.0710 | 0.013 | -5.629 | 0.000*** | -0.096 | -0.046 |
| **complaints about sliders** | -0.0781 | 0.032 | -2.466 | 0.014* | -0.140 | -0.016 |
| **complaints about sliders other worker** | 0.0038 | 0.030 | 0.126 | 0.900 | -0.055 | 0.062 |

Table 4.5.: OLS Regression on tip given to workers

*Notes:*
[1] Standard Errors are robust to cluster correlation (cluster on supervisor level)
[2]$p < 0.05^*$; $p < 0.01^{**}$; $p < 0.001^{***}$
[3] Categorical variables are labelled as "C" and compare against the baseline categories: Risky-TP (T10) for treatment, Female for Sex, and Republican for politics. Own party worker and Worker wins tournament are dummy variable with yes coded as 1.

## 4.4. Discussion and Conclusion

In this study, we have used a novel adaptation of the slider task to build a stylised performance review task and combine it with a simple generosity task. With these tasks, we measure social preferences of Prolific participants based on their U.S. party affiliation. The results on social preferences within the performance review task can be fairly summarised as a null-result. But at the same time, we show that within the task to give an additional 'tip', the participants show significant and substantial in-group favouritism.

Reflecting on the quantitative results on social preferences, we gain several important methodological insights on how this set of Prolific participants dealt with this task. First of all, participants take a given task seriously, regardless of whether they are additionally incentivised for it or not: Even the supervisors with a fix bonus payment who have no financial reason to engage with the performance review task at all, still do so. If they would do nothing instead, that would be to the disadvantage of the workers. The supervisors with risky bonus payments engage with the task in a way that maximises their own bonus payment. Only when the task directly prompts them to show social preferences (i.e., the opportunity to give a tip) and has no effect on their bonus payment, the supervisors show clear social preferences as in-group favouritism. This has implications for research on social preferences on Prolific. Based on the behaviour within this study and the contrasting results in its two tasks, we can derive a hierarchy of what participants prioritise when engaging in this study. **(1.)** Maximising their own bonus payment has first priority, as we see through the supervisors with risky bonus payments. **(2.)** Consciously engaging with the task. Supervisors fulfil tasks even when they give themselves no added financial bonus, such as giving tip or engaging in the performance review task. This can be conceptualised through the model of a 'moral horizon' described in chapter 2, where participants in an experiment commit to the values and expectations within that experimental moral horizon. **(3.)** Showing their social preferences. Only third, when no own bonus payment is involved, and the task itself is open to show social preferences, then the participants in this study do so.

This brings the question to broader research on social preferences on Prolific. When it seems that showing their social preferences is by far not the first priority for participants, then how does a task need to be structured to elicit them? And how much demand effect would that entail? Further research should ask these methodological questions to investigate whether we can confirm certain hierarchies of values within Prolific participants and whether clusters of Participants can be detected. Such clusters might be, but do not necessarily have to be correlated with demographic information easily accessible on Prolific or personality measures such as Big-Five. A first step to approach this research would be a meta-analysis followed by targeted studies to find evidence or falsify the resulting hypotheses.

The present study naturally carries features that limit the generalisibility of the results. Within the performance review task, the risk-neutral *sat* is dictated by the deduction of points when a complaint occurs. The amount of point deduction could be varied to see how this influences behaviour. Furthermore, the operationalisation of identity by focusing on U.S. party affiliation is only one application. Other application in different cultures might lead to different results.

Given the limitations, this study presents an approach to measure social preferences stylised as a performance review, combines it with a more classic generosity task of giving tips, and produces methodological insights on how participants handle and engage in a study. To produce valuable social research, further research should take upon the presented methodological questions and continue to validate and benchmark different measures and strengths of intervention on the same population.

# Appendix

## 4.5. Result (I): Heterogeneous behaviour types remain largely stable

The following analysis aims to identify behaviour types beyond average treatment effects. To identify clusters of individuals who show similar behaviour, we use the Mean Shift algorithm, which is an unsupervised machine learning method. Mean Shift is a centroid-based algorithm that updates the centre of each cluster to be the average (mean) of the points within a specified region. One key advantage of Mean Shift is that it does not require us to specify the number of clusters in advance. This makes it flexible and effective at finding clusters based on the natural structure of the data. For clarity in our visualisations, we only consider clusters that contain at least three individuals.

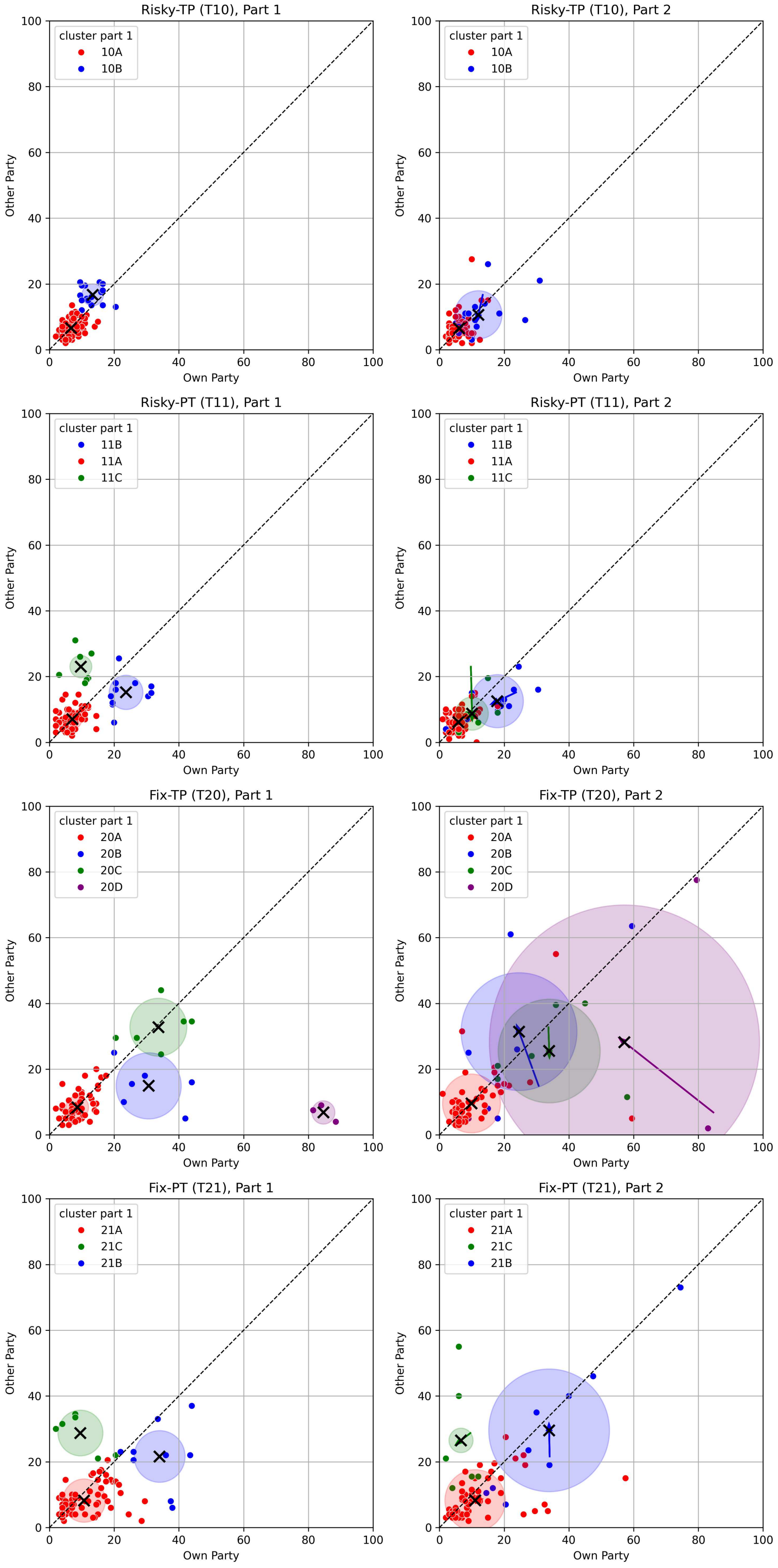


Figure 4.7.: Behaviour Types

The results are presented in figure 4.7 in a series of scatter plots that show the *sat*-scores a supervisor gave to the own party worker versus the other party worker. The dashed 45°-line represents an equal *sat*-score for both workers. A dot below the 45°-line represents a supervisor who was more lenient with the own-party worker because the threshold of accepting sliders is higher for the own party worker than the other worker. In reverse, a dot above the 45°-line represents a worker who was more lenient towards the other party worker.

Each row represents a different treatment, ordered as Treatment *Risky-TP (T10)*, *Risky-PT (T11)*, *Fix-TP (T20)* and *Fix-PT (T21)*. For each treatment, two plots are shown: Part 1 on the left and Part 2 on the right.

Each scatter plot uses colours to represent various clusters. Red represents cluster A, the cluster most close to the origin. Blue represents cluster B, the second largest cluster, which in all treatments except for treatment *Risky-TP (T10)* favours the own party worker. Green represents cluster C, and purple represents cluster D. The clusters are built upon the data in part 1. We are interested in whether behavioural types also show similar behaviour in part 2. Therefore, we plot each individual in part 2 according to their behaviour in part 2 and keep the assignment of cluster and colour constant. This allows us to detect changes in cluster behaviour. In the part 1 plots, an "x" additionally marks the mean positions of each cluster, and ellipses around these mean points depict the standard deviations. In the part 2 plots, the mean positions ("x") and standard deviations for part 2 are shown, along with a vector that indicates the movement from the part 1 to part 2 mean positions. These vectors are marked in solid lines leading to the "x," representing the mean of the cluster in part 2. To determine that a cluster has made a substantial change in behaviour, we evaluate whether the mean position in part 1 (i.e., the start of the vector in part 2) is outside of the standard deviation of the cluster in part 2. [2]

This visualisation highlights the shifts in *sat*-scores from part 1 to part 2, illustrating how each cluster's score changes. The ellipses provide a sense of the variability within each cluster, and the vectors clearly show the direction and magnitude of the shifts between the two parts. Interestingly, in both treatments with risky bonus payment, all clusters are close to the 45°-line and the origin and have a small standard deviation. The only cluster where the mean of part 1 is not within the standard deviation ellipse around the mean of part 2 is cluster 11C. It moves even closer to the 45°-line and explains part of the treatment averages of *Risky-PT (T11)* in table 4.4. Cluster 20D has a very large standard deviation in part 2. This is due to only three observations in part 1 that behave differently in part 2, which should be interpreted as outlier behaviour. All other clusters show similar behaviour in part 2 as in part 1.

As a limitation to this approach it has to be mentioned that different clustering algorithms might lead to slightly different results.

[2] Evaluating whether the mean in part 2 is outside of the standard deviation in part 1 would lead to more "false" positives: The assignment of clusters in part 1 is made with the purpose to minimise the standard deviation in part 1. Hence, by design small changes of behaviour would lay outside. And these small changes in behaviour are to be expected to a certain degree.

## 4.6. Construction Slider-Acceptance-Threshold (sat)

As a central variable we measure the ‘slider-delta’ i.e. the distance a worker has placed the slider from the exact middle of the range. The main dependent variable is whether a supervisor accepts or rejects a slider. Out of this we calculate the acceptance-threshold of ‘slider-delta’ from which onwards a supervisor accepted sliders for a specific worker. We do not expect this to be a clear point, but some noisy behaviour. To construct this measure, we will (1) build the percentile-rank of every accepted slider over their delta. (2) then the percentile-rank of all rejected sliders and take the inverse (100 – pct). (3) Look at which slider-delta these two numbers are equal. This is the ’slider-acceptance-threshold’ (’sat’) for a supervisor for this specific worker.

# Part II.

# Extending The Experimental Framework

# 5. Motivated Sampling Of Information

## Authors

Leon Houf & Vinicius Ferraz [1]

## Abstract

This paper investigates information sampling in a situation where objective and subjective criteria are coupled. This creates a situation that gives room for motivated reasoning, which we identify as motivated sampling. We present participants with a binary sampling and decision task. Participants sample information from two “computers,” which generate numbers from distinct distributions, and participants have to identify the ‘high distribution’ computer. In this task, we vary externalities on the participants’ decision to induce subjective preferences. Furthermore, we vary the type of feedback participants receive. We find motivated reasoning in several instances. First, we show that female subjects sample significantly more than male subjects when faced with a negative externality or Bayesian posterior feedback. Furthermore, we show a strong intensive margin of motivated sampling. Here, subjects sample additionally from the option with positive externality if they deem it correct which shows an added liking to sample from it. These findings provide an understanding of motivated sampling as a potential channel leading to confirmation bias. It emphasises the importance of subjective preferences, feedback and gender differences in all situations where information sampling is necessary for decision-making.

## Keywords

Motivated Reasoning, Information Sampling, Bayesian Learning, Decision Theory

## 5.1. Introduction

Decision-making is an integral part of human life. Individuals are frequently faced with the task of selecting from multiple options, whether it’s choosing a restaurant, booking travel tickets, or picking a university. In these situations, information sampling is crucial to facilitate well-informed choices. Understanding the dynamics of information sampling can lead to better decision-making, especially in complex scenarios. Previous research has provided valuable insights into information sampling behaviour and its cognitive and computational costs for human subjects (Kool & Botvinick, 2018; Petitet, Attaallah, Manohar, & Husain, 2021), individual factors in sampling and information-seeking (Gottlieb & Oudeyer, 2018; Kelly, Sharot, et al., 2021), decision-making perspectives (Leung, 2020) and how rewards influence sampling through a Pavlovian-approach (Hunt, Rutledge, Malalasekera, Kennerley, & Dolan, 2016).

[1] **Status:** This paper was presented at the BSE Summer Forum in Barcelona in June 2023 and at SPUDM in Vienna in August 2023. Submission to a journal in the field of judgement and decision making is in preparation.

However, there remain open questions about the coupling of subjective and objective criteria in decision-making. Specifically, how do individuals approach decisions when they aim for the objectively best option but already have a pre-existing preference? This phenomenon, known as *motivated reasoning*, is characterised by individuals processing information in a way that aligns with their pre-existing beliefs or desires (Bénabou & Tirole, 2016; Eil & Rao, 2011; Hagenbach & Koessler, 2022). In this paper we introduce the concept of *motivated sampling.* This phenomenon represents the overlap between information sampling and the well-established idea of motivated reasoning. Furthermore, it is a potential channel explaining confirmation bias, building on the 'positive confirmation bias in the acquisition of information' described by Jones and Sugden (2001).

To disentangle the effects of objective and subjective criteria in information sampling, we present subjects with a binary decision task. In this task, subjects have to sample information to determine the objectively correct option to receive a payoff. We then asymmetrically add negative and positive externalities to the options. A positive externality is an additional reward for an organisation the subject liked, while a negative externality is a reward for an organisation the subject explicitly disliked. Through this, we induce subjective preferences into the sampling and decision situation. Using a between-subjects design, we can measure how these subjective criteria affect sampling behaviour. Our central research question is: 'How do subjective preferences on externalities influence motivated sampling?' Additionally, we analyse the accuracy of posterior beliefs with different forms of feedback and the time participants actively engage in the task.

Our findings show that women sample significantly more information in total than men when a negative externality is at play. Subjects sample much more for the option with a positive externality when they deem this option correct than when incorrect. This behaviour, termed motivated sampling, indicates a 'liking' to sample from options that meet not only objective criteria but also subjective ones. In contrast, when a negative externality is involved, we do not see such a behaviour. For both types of externality, male participants show a stronger bias for the "nicer" option than female participants.

We offer a novel perspective on information sampling strategies by disentangling the effect of objective and subjective criteria. Specifically, we uncover the mechanisms of motivated sampling. Hereby, we add a specific application to the more broadly defined theme of motivated reasoning. This also serves as a fundamental, underlying mechanism of confirmation bias.

The remainder of this paper is structured as follows: We first outline the method and experimental design. Then we present the empirical results of our experiment. Section four concludes.

## 5.2. Method

This section provides an overview of the experimental design used in our online study.

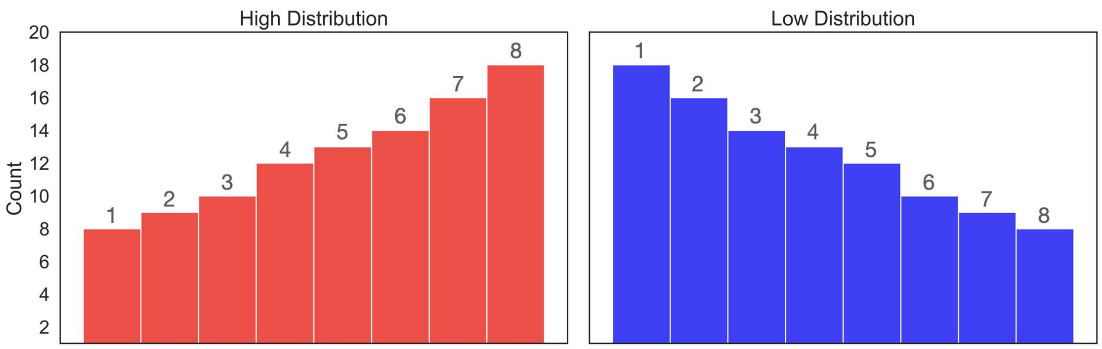


Figure 5.1.: High and Low Distribution

### Demographics

A total of 457 students were recruited from the experimental economics labs at Heidelberg University and Rhine-Waal University of Applied Sciences, with a 37% representation from Heidelberg and 63% from Rhine-Waal. The participants were split almost equally between male (48%) and female (51%) with 1% choosing not to disclose or identifying as non-binary. The average age of the participants was 24.2 years old (standard deviation 4.6 years) and they represented a diverse mix of nationalities, with the largest group being German (48%) followed by Indian (10%). The remaining participants came from a variety of international backgrounds. The experiment was programmed using oTree (D. L. Chen et al., 2016). The experiment was online and participants could take part any time of the day or week and take breaks of any length. The median participant had the experiment open in the browser for 30.5 minutes. The average earning was 4.91€.

### Description of Main Task

During the task, participants are presented with two “computers”. Each computer generates numbers based on specific distributions, as depicted in figure 5.1. One of the computers produces numbers that are higher on average because the computer uses a ‘high distribution’ of numbers, whereas the other computer uses a ‘low distribution’. Both distributions produce numbers from 1 to 8, as used in Goette, Han, and Leung (2020). The ‘high computer’ produces numbers using the distribution shown on the left side of figure 5.1 and the ‘low computer’ produces numbers using the distribution on the right. In every round, one computer uses the high distribution of numbers, while the other computer uses the low distribution. The computer that uses the high distribution is determined randomly in each round with a 50/50 chance. The participants’ goal is to identify which computer uses the high distribution. Each correct identification is rewarded with a point. Participants can sample as many numbers as they want by clicking on one of the computers, but they are restricted to only sample one new number every two seconds. This is explained as a need of the computers to reload to produce the next number. We use this two-second restriction to prevent rapid “over-clicking” by participants because we want to create a situation where every new information can be taken into account subsequently.

### General Procedure of Experiment

Now, we describe the overall procedure of the experiment surrounding the main task. Before the experiment starts, we gather demographic information from participants. This includes two questions about which organisations participants would be *most* and *least* likely to contribute money to. Then, the experiment starts with a practice round and 20 payment relevant rounds. Each round consists of three parts: Part A assesses the prior belief about which computer uses the high distribution, shown in figure 5.7 and 5.8. Part B is the main task explained above. Part C assesses the posterior belief after the participants made their selection, shown in figure 5.9. At the end of the 20 rounds, we randomly select three rounds as payoff relevant. For each point that a participant scored in those rounds, they receive 1.5€, in addition to a 1.5€ show-up fee.

We have nine treatments that differ in whether an externality is added to the main task and whether feedback is provided after part C as the end of a round. Subjects are randomly assigned to their treatment group. The treatments are created as a 3x3 design along the dimensions of externality and feedback. And as pre-registered[2], we have 50-52 participants in every treatment cell as shown in table 5.1, allowing us to pool the cells across rows or columns when we compare the respective treatment dimension for feedback and externality.

| | | Feedback | | |
|---|---|---|---|---|
| | Condition | Outcome | Bayes | No |
| | Negative | 51 | 50 | 52 |
| Exter-nality | No | 52 | 51 | 50 |
| | Positive | 50 | 51 | 50 |

Table 5.1.: Subjects per Treatment

### Externality Treatments

In the externality dimension we distinguish between no externality, positive externality and negative externality. With no externality, the main task is exactly as described above. In the externality treatments, we attach an externality randomly to one of the two computers in each round.

In the positive externality treatments, we use the organisation the participant chose in the demographics section as the organisation they are *most* likely to give money to. This organisation is attached to one of the computers, and also the organisation receives 1 point in the round if the participant chooses the respective computer and it is *correct* (shown in appendix figure 5.11). This introduces a subjective element to the decision task. While the objective criteria is still to select the correct computer, since only then the participant *and* the preferred organisation receive a point, participants might have a subjective preference to experience the option with externality to be correct.

In the negative externality treatments, we use the organisation the participant chose as *least* likely to give money to. This organisation is attached to one of the computers, and

[2] AsPredicted #104844

the organisation receives 1 point in the round if the participant chooses the respective computer but that was the *wrong* decision (shown in appendix figure 5.12). Here, the objective incentive is still to select the correct computer, since then the participants receive a point and the antagonising organisation receives nothing. The subjective element for the participants here is an increased subjective incentive to not be wrong when selecting the option with externality as the antagonising organisation can only receive a point through a subject's mistake.

### Feedback Treatments

In the feedback treatment dimension, we distinguish between no feedback, outcome feedback and Bayes feedback. With no feedback, the procedure is exactly as described above and participants move to the next round without receiving any feedback.

In the outcome feedback treatments, participants learn at the end of each round whether their computer choice was correct or incorrect.

In the Bayes feedback treatments, at the end of each round after stating their posterior belief in part C, participants receive a reminder of the posterior belief they just stated and are informed about the rational Bayesian posterior. The feedback is shown in appendix figure 5.14 and the calculation of the Bayesian posterior is outlined in appendix 5.5.1. The overall design is summarised in figure 5.2.

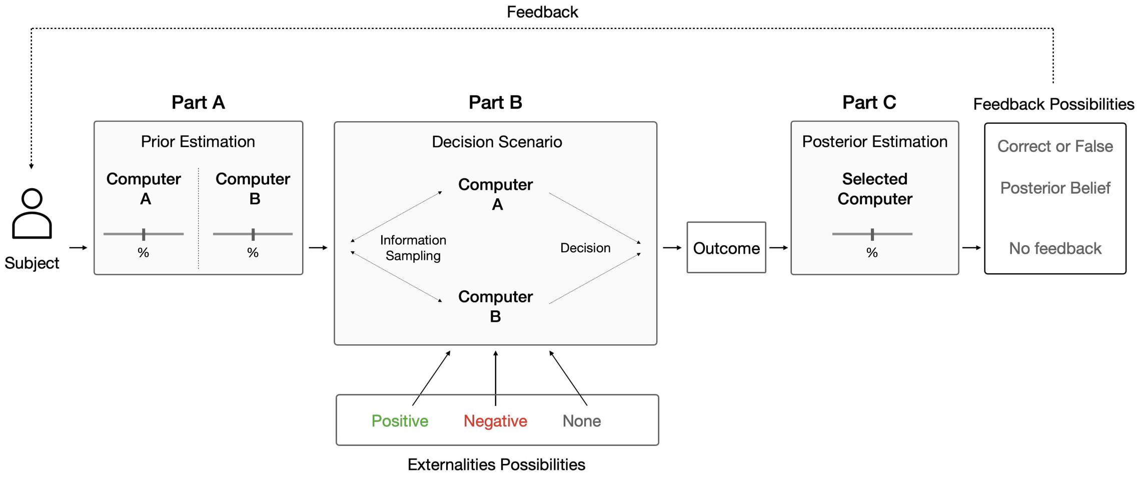


Figure 5.2.: Summary Experimental Design

## 5.3. Experimental Results

This section presents the results of the human subject experiment. First, we see the results for total sampling $S$ per round across the treatment dimensions and how these results are driven by gender. We continue with the analysis of motivated sampling. First, we analyse the the extensive margin of unequal sampling between the two available options $A$ and $B$. This we will split by whether the externality option $A_{ext}$ was selected or not. We then analyse the intensive margin of this unequal sampling, also split by whether or not the

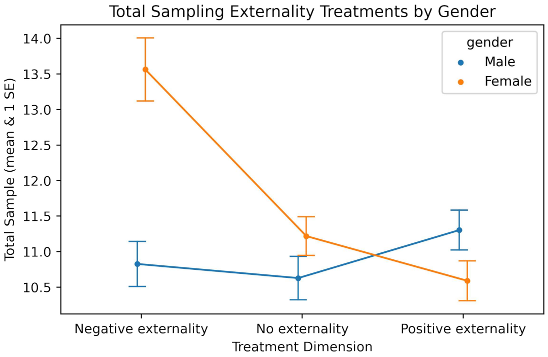


Figure 5.3.: Sampling by gender: externality treatments

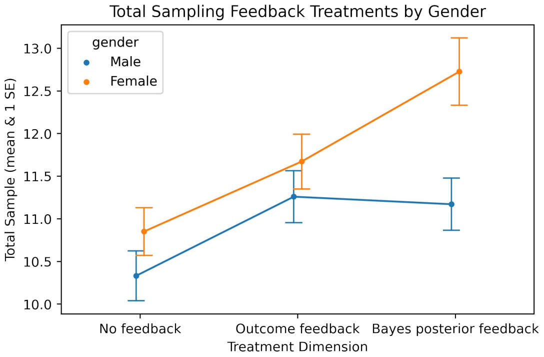


Figure 5.4.: Sampling by gender: feedback treatments

externality option $A_{ext}$ was selected. Here we create our measure of the intensive margin of motivated sampling $MotSamp$. Furthermore, we analyse the decision behaviour and scoring success of subjects. We measure the time subjects actually take for sampling. Lastly, we show the accuracy of the stated posterior belief.

### 5.3.1. Total Sampling per Round

First, we analyse the total sampling behaviour $S$ per round. Here we will look at the effect of the externalities and feedback treatments on the total number of samples a subject created for both computers together in one round.

We find that participants sample most in the negative externality (12.18 samples per round), and equally in the no- and positive externality treatments (11.01 and 10.95). Across the feedback dimension, participants sample more, the more detailed feedback they receive: Bayes feedback (12.00) $>$ outcome feedback (11.54) $>$ no feedback (10.61).

Figures 5.3 and 5.4 show that these effects are driven by gender. Female participants sample significantly more than male participants in the negative externality treatments and in the Bayes feedback treatments. This shows that women make higher sampling effort when the context is most salient: either through a negative externality or detailed feedback on their own stated posterior belief.

### 5.3.2. Motivated sampling

After this effect of total sampling per round, we turn to motivated sampling within a round. We define motivated sampling as the tendency to sample additionally *because* of subjective preferences. To identify this behaviour we need to perform an analysis in multiple steps.

First, we will identify the extensive margin of whether subjects sampled unequally for the two available options. Then we will split this by whether the externality option ($A_{ext}$) was selected or the non-externality option ($B_{non}$). This will give us a score of the extensive margin of motivated sampling, but more importantly, we use it to then move to the intensive margin of how much subjects sample more when they sample unequally. We will then also split by whether the externality option $A_{ext}$ was selected or not, which gives the crucial comparison of the sampling behaviour when subjects deem $A_{ext}$ correct compared to when they deem $B_{non}$ correct. This allows us to identify the additional sampling from an option out of a subjective preference for doing so. Out of this, we will calculate this score of motivated sampling in a round $MotSamp$, split by type of externality and gender.

### Extensive Margin of Unequal Samples

Many subjects might use a strategy where they always sample equally from both options $A$ and $B$. Those subjects will not show a behaviour of motivated sampling within a round. Yet many subjects might sample unequally from the two options, resulting in unequal sample sizes $s_A \neq s_B$. We will first turn to the extensive margin of this unequal sampling behaviour, of whether a subject did sample unequally or not.

We measure for each participant in how many of the 20 rounds they show an unequal sample strategy $s_A \neq s_B$. Figure 5.5 plots this fraction of rounds in which a subject showed an unequal sample. 23% subjects show a fraction of 0, so never sampling unequally, i.e. always sample the same number of information from both sources. In total 49.7% of all subjects show unequal samples in only a quarter of the rounds or less. This is drastically more than compared to 6.8% who always sample unequally and 25.6% who show unequal samples in at least three-quarters of all rounds. This pattern is stable over all treatments.

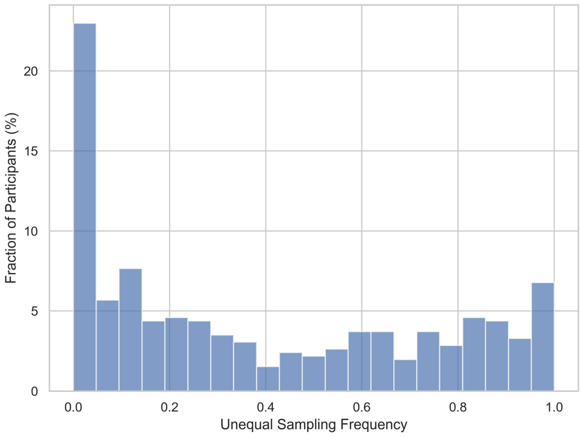


Figure 5.5.: Unequal sampling by subjects

### Extensive Margin of Unequal Samples with Externality Selection

Table 5.2 shows the extensive margin of how many observations in a treatment exhibit an unequal sample. In the first column, overall by treatment, where treatments with an externality show more unequal samples and especially the negative externality treatment shows statistically significant more unequal sampling than with positive externality ($p < 0.01$) and without externality ($p < 0.001$) using a Chi-Squared Test. Columns two and three split this overall value in whether the externality or non-externality option was selected. Now, we can calculate the extensive margin of motivated sampling.

With a negative externality, we see $42.7 - 40.3 = 2.4$ percentage points more unequal samples when the externality is selected. With a positive externality, we observe $38.3 - 37.9 = 0.4$ percentage points more unequal samples, so in both cases a rather mild extensive margin of motivated sampling.

| Externality | Overall | Ext select | Non ext select |
|---|---|---|---|
| No | 37.1% | - | - |
| Negative | 41.5% | 42.7% | 40.3% |
| Positive | 38.1% | 38.3% | 37.9% |

Table 5.2.: Extensive Margin of Unequal Sampling

**Intensive Margin of Motivated Sampling**

Now, we turn to the intensive margin, so how many more subjects sample for one of the options when they showed a $s_A \neq s_B$ unequal sampling. For this, we calculate the $\Delta$-sample, $\Delta s$, which is the difference between the sample for the selected option, $s_{select}$, and the option that was not selected, $s_{NonSelect}$, $\Delta s = s_{select} - s_{NonSelect}$.

Table 5.3 shows $\Delta s$ in the first column by treatment. Here, we see motivated sampling as the intensive margin with no externality is with $\Delta s = 0.419$ sample on average lower than both intensive margins with negative (0.876) and positive (0.912) externality.

Columns two and three split $\Delta s$ by whether the option with externality was selected, $\Delta s_{ext}$, or without, $\Delta s_{non}$.

| Externality | $\Delta s$ | $\Delta s_{ext}$ | $\Delta s_{non}$ |
|---|---|---|---|
| No | 0.419 | - | - |
| Negative | 0.876 | 0.842 | 0.91 |
| Positive | 0.912 | 1.769 | -0.075 |

Table 5.3.: Intensive Margin of Unequal Sampling $\Delta s$

With $\Delta s_{ext}$ and $\Delta s_{non}$ we can now calculate the intensive margin of motivated sampling as the additional sampling when the externality option $A_{ext}$ was selected as $MotSamp = \Delta s_{ext} - \Delta s_{non}$. With a negative externality, we see almost no motivated sampling within a round ($MotSamp = 0.842 - 0.91 = -0.068$) as the margin of unequal sampling is relatively similar regardless of whether subjects select the externality or not.

Yet with a positive externality, there is a striking difference. Here we see a very strong case of motivated sampling of $MotSamp = 1.769 - (-0.075) = 1.844$. This shows that subjects sample much more additionally from the positive externality option, hence "like" sampling from it when they deem it correct.

Table 5.4 splits this analysis of motivated sampling by gender. With both types of externalities, we observe that male participants show a stronger liking for the "nicer" option than female participants, i.e. the option with positive externality or the option *without* negative externality, respectively.

| Externality | Overall | Male | Female |
|---|---|---|---|
| Negative | -0.068 | -0.201 | 0.200 |
| Positive | 1.844 | 1.925 | 1.637 |

Table 5.4.: Motivated sampling $MotSamp$

### 5.3.3. Decision and Scoring Behaviour

Next to the sampling behaviour, we are also interested in how the decision and scoring behaviour are influenced by externalities and feedback. Here we should note that the correct option was always determined randomly, so the correct baseline is always a 50/50 split. And indeed we see that the option that subjects select is relatively equally balanced between the option with and without externality with a negative externality (49.0% to 51.0%). With a positive externality subjects decide in 53.3% for the externality option. This is significantly more than the balanced split ($p < 0.001$) in a binomial test, as seen in table 5.5, mirroring the motivated sampling into decision behaviour.

| Externality | Externality Select | Non-ext select |
|---|---|---|
| Negative | 49.0% | 51.0% |
| Positive | 53.3% | 46.7% |

Table 5.5.: Select Ext / Non-Ext per Externality Treatment Dimension

Interestingly, this does not translate into meaningful differences in correct decisions, as all treatments hover around 75% accuracy in their decisions, see table 5.6. This is also very stable over the course of the experiment of 20 rounds.

| Treatment | Scoring |
|---|---|
| OVERALL | 75.2% |
| FEEDBACK | |
| Outcome | 74.5% |
| Bayes | 77.0% |
| No | 74.2% |
| EXTERNALITY | |
| Negative | 75.3% |
| Positive | 74.8% |
| No | 75.5% |

Table 5.6.: Scoring across treatments

### 5.3.4. Page Time Analysis

Furthermore, as an exploratory analysis, we investigate the time subjects take to complete the core sampling task. Figure 5.6 reports on the x-axis the seconds subjects spend on the page for the main task from the start of sampling till confirmation of their decision. We plot the $5^{th}$, $25^{th}$, $50^{th}$, $75^{th}$ and $95^{th}$ percentile and the mean of the seconds they spent on the page for the task. An interesting finding for further research is that even though subjects in the negative externality treatments sample more, they spend less time on the task than subjects in the other treatments. Potential explanations could be that through more sampling they take less time to evaluate the samples, or that the pure unpleasant presence of the negative externality prompts subjects to be faster in their evaluation process, which would show another form of motivated effort distribution, a hypothesis for further research.

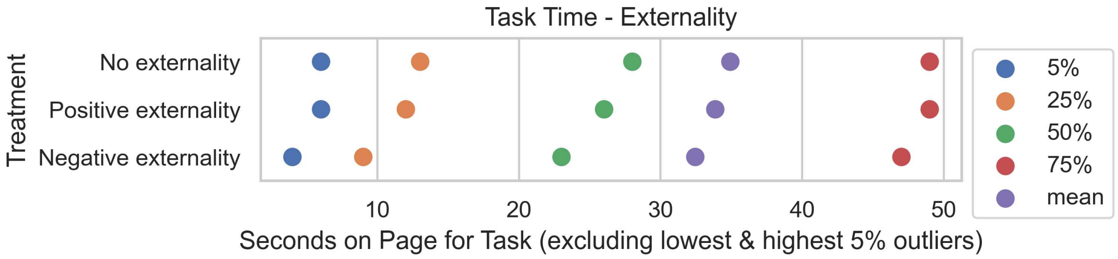


Figure 5.6.: Task time

### 5.3.5. Posterior Beliefs

After subjects completed the task as part B of a round, we ask them “Please let us know how likely it seems to you that your choice was correct”. Table 5.7 reports the percentage point difference between their stated belief and the rational Bayesian posterior. We see that subjects on average gave a lower estimation than the rational Bayesian posterior, where the subjects who receive the Bayes feedback are closest to the rational posterior since they get feedback on their stated posterior and the Bayesian posterior in every round and can learn from it. Interestingly, subjects with only outcome feedback have a significantly higher difference than subjects with no feedback. As preregistered, we excluded participants who stated a posterior belief lower than 50%, as a potential sign they did not seriously think about the question.

| Treatment | Difference |
|---|---|
| Overall | -4.4% |
| FEEDBACK | |
| Outcome | -6.7% |
| No | -4.2% |
| Bayes | -2.3% |

Table 5.7.: Difference Stated- to Bayesian-Posterior

## 5.4. Discussion & Conclusion

In this study, we have investigated the intersection of information sampling and motivated reasoning, which we term motivated sampling.

We find substantial evidence for motivated sampling, which emerges in different forms. First, we observe that female participants sample significantly more in context-rich environments, especially when the externality is negative. Furthermore, we find that subjects show a strong sense of motivated sampling when they deem the option with a positive subjective preference correct. Here, they show a behaviour of "liking" to sample from it additionally. This behaviour of motivated sampling is stronger pronounced for male than for female subjects. This translates into a tendency to also select the option that is associated with the positive externality more often than it would be objectively correct. As complementary findings, we observe that subjects use the least time for the task in the negative externality treatment, even though they sample the most. If subjects receive feedback on the rational Bayesian posterior, their stated posterior belief is very close to it. When subjects receive feedback on the outcome of their decision, the stated posterior belief is more distant from the Bayesian posterior than when subjects receive no feedback at all.

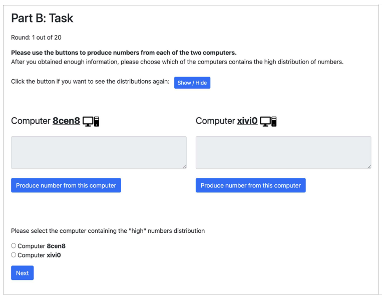
With this study and its findings, we contribute to a better understanding of information sampling behaviour as a specific application of the broader field motivated reasoning. This indicates different cost functions of sampling depending on externality, relating to Petitet et al. (2021) and Kelly et al. (2021).

Our study shows that the amount and direction of effort spent on information sampling in a decision-making context is greatly influenced by subjective criteria. The gender-based differences, particularly among female participants in specific treatments, indicate that context might influence information acquisition differently across genders, potentially having implications in areas like marketing, education, or policy-making.

The study naturally carries several limitations. The study's online nature might introduce biases, as participants could be influenced by external factors while doing the experiment not present in controlled lab settings. The list of organisations provided as externality is inherently incomplete to elicit strong subjective preferences for all subjects. It is thereby expected that some participants had no "strong feelings" about any of them, which reduces the studied effect of externalities. Furthermore, experienced subjects might not have been truthful in their response to the questions about the organisations, anticipating that they might play a role later and therefore just indicating organisations they feel indifferent about. This would imply that our findings provide a lower bound of the effect of externalities in motivated sampling.

Future studies could explore the underlying psychological factors driving the observed behaviours, especially in negative externality scenarios. It would also be intriguing to analyse further the gender differences and what might drive them. Additionally, expanding the study to diverse demographic groups or introducing more complex decision-making tasks could provide richer insights. The behaviour termed motivated sampling can furthermore be seen as a fundamental, underlying mechanism of confirmation bias, which future studies should explore specifically.

This research highlights the role of objective and subjective criteria in information sam-



Part B: Task

Round: 1 out of 20

**Please use the buttons to produce numbers from each of the two computers.**
After you obtained enough information, please choose which of the computers contains the high distribution of numbers.

Click the button if you want to see the distributions again:

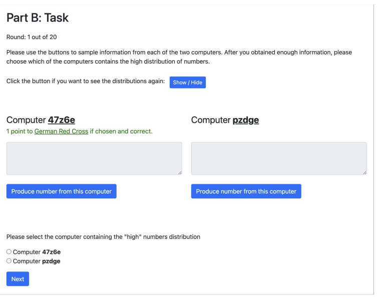
Computer **8cen8**

Computer **xivi0**

Please select the computer containing the "high" numbers distribution

○ Computer **8cen8**
○ Computer **xivi0**

Show / Hide
Produce number from this computer
Produce number from this computer
Next



Part B: Task

Round: 1 out of 20

Please use the buttons to sample information from each of the two computers. After you obtained enough information, please choose which of the computers contains the high distribution of numbers.

Click the button if you want to see the distributions again:

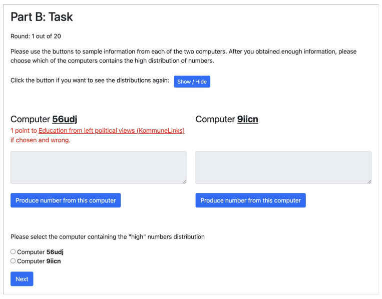
Computer **47z6e**
1 point to German Red Cross if chosen and correct.

Computer **pzdge**

Please select the computer containing the "high" numbers distribution

○ Computer **47z6e**
○ Computer **pzdge**

Show / Hide
Produce number from this computer
Produce number from this computer
Next



Part B: Task

Round: 1 out of 20

Please use the buttons to sample information from each of the two computers. After you obtained enough information, please choose which of the computers contains the high distribution of numbers.

Click the button if you want to see the distributions again:

Computer **56udj**
1 point to Education from left political views (KommuneLinks) if chosen and wrong.

Computer **9iicn**

Please select the computer containing the "high" numbers distribution

○ Computer **56udj**
○ Computer **9iicn**

Show / Hide
Produce number from this computer
Produce number from this computer
Next

Figure 5.10.: No externalities Figure 5.11.: Positive extern. Figure 5.12.: Negative extern.

pling. The findings, particularly regarding gender differences and the influence of personal values, underscore the importance of the phenomenon of motivated sampling.

# Appendix

## 5.5. Experimental Design Details

### 5.5.1. Calculation of Bayesian Posterior

Assumptions:

- prior information that each computer produces number from one of the two (high and low) distributions is given,

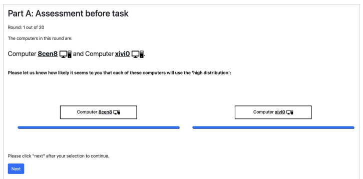
- before any observations, each computer is equally likely to be sampling from high as from low (prior = 50/50).

Here $HL$ indicates that computer 1 contains samples from the high and computer 2 contains samples from the low distribution, and $X$ correspond to the array of all sampled numbers from both computers. Transferring it to the Bayes' theorem we have:

$$P(HL \mid X) = \frac{P(X \mid HL)P(HL)}{P(X \mid HL)P(HL) + P(X \mid LH)P(LH)} \tag{5.1}$$

$P(HL \mid X)$ is the probability that computer 1 is high and computer 2 is low.

### 5.5.2. Prior and Posterior



Part A: Assessment before task

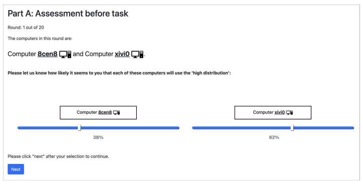
Round: 1 out of 20

The computers in this round are:

Computer **8cen8** and Computer **xivi0**.

**Please let us know how likely it seems to you that each of these computers will use the 'high distribution':**

Please click "next" after your selection to continue.



Part A: Assessment before task

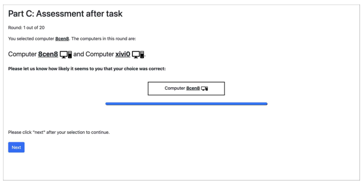
Round: 1 out of 20

The computers in this round are:

Computer **8cen8** and Computer **xivi0**.

**Please let us know how likely it seems to you that each of these computers will use the 'high distribution':**

Please click "next" after your selection to continue.



Part C: Assessment after task

Round: 1 out of 20

You selected computer 8cen8. The computers in this round are:

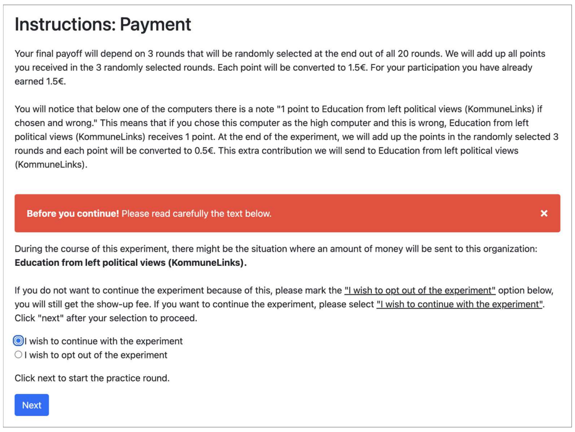
Computer **8cen8** and Computer **xivi0**.

**Please let us know how likely it seems to you that your choice was correct:**

Please click "next" after your selection to continue.

Figure 5.7.: Before click Figure 5.8.: Prior after click Figure 5.9.: Posterior

### 5.5.3. Externalities



**Instructions: Payment**

Your final payoff will depend on 3 rounds that will be randomly selected at the end out of all 20 rounds. We will add up all points you received in the 3 randomly selected rounds. Each point will be converted to 1.5€. For your participation you have already earned 1.5€.

You will notice that below one of the computers there is a note "1 point to Education from left political views (KommuneLinks) if chosen and wrong." This means that if you chose this computer as the high computer and this is wrong, Education from left political views (KommuneLinks) receives 1 point. At the end of the experiment, we will add up the points in the randomly selected 3 rounds and each point will be converted to 0.5€. This extra contribution we will send to Education from left political views (KommuneLinks).

**Before you continue!** Please read carefully the text below. ×

During the course of this experiment, there might be the situation where an amount of money will be sent to this organization: **Education from left political views (KommuneLinks).**

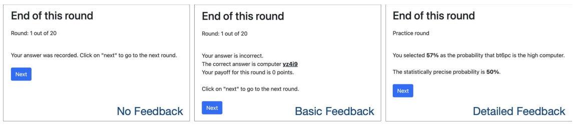
If you do not want to continue the experiment because of this, please mark the "I wish to opt out of the experiment" option below, you will still get the show-up fee. If you want to continue the experiment, please select "I wish to continue with the experiment". Click "next" after your selection to proceed.

● I wish to continue with the experiment
○ I wish to opt out of the experiment

Click next to start the practice round.

Figure 5.13.: Opt out in case of negative externalities

### 5.5.4. Feedback



**End of this round**

Round: 1 out of 20

Your answer was recorded. Click on "next" to go to the next round.

**End of this round**

Round: 1 out of 20

Your answer is incorrect.
The correct answer is computer **yz4i9**
Your payoff for this round is 0 points.

Click on "next" to go to the next round.

**End of this round**

Practice round

You selected **57%** as the probability that bt6pc is the high computer.

The statistically precise probability is **50%**.

Figure 5.14.: Feedback

# 6. Trust in the Machine: How Contextual Factors and Personality Traits Shape Algorithm Aversion and Collaboration

## Authors

Leon Houf, Vinícius Ferraz, Thomas Pitz, Christiane Schwieren & Jörn Sickmann [1]

## Abstract

This paper studies the effects of contextual factors and personal variables on algorithm aversion in decision delegation. An experimental design with four treatments—baseline, explanation, payment, and automation—was used to examine subjects' choices to delegate decisions to an algorithm with hidden expected values. We evaluated the impact of Big Five personality traits, locus of control, generalised trust, and demographics alongside the treatment effects using statistical analyses and machine learning models, including Random Forests, Gradient Boosting Machines, and Uplift Random Forests. Results show that payment reduces delegation, whereas full automation increases it. Age, extraversion, openness, neuroticism, and locus of control significantly predicted delegation behaviour. Additionally, female participants reacted more strongly to algorithm errors. Increased delegation rates improved algorithm accuracy. These findings provide new insights into the roles of contextual conditions, personal variables, and gender in shaping algorithm aversion, offering practical implications for designing user-centric AI systems.

## Keywords

Algorithm aversion, Human-computer Interaction, Decision Behaviour, Machine Learning, Causal Inference

## 6.1. Introduction

Driven by technological advancements, data availability, and computing power, intelligent systems powered by Artificial Intelligence (AI) have become common in our society due to their transformative potential (Russell, 2010). AI simulates human behaviours like learning and decision-making (McCarthy, 2007). AI's ability to efficiently process vast amounts of data, inform decisions, and automate processes has led to widespread adoption (Azucar, Marengo, & Settanni, 2018). However, these technological shifts can lead to new social phenomena, such as algorithm aversion, characterised by the reluctance to use algorithms in decision-making despite their superior ability to undertake specific tasks (Dietvorst,

[1] **Status:** This paper was presented at the BSE Summer Forum in Barcelona in June 2023, at the Workshop on Gender in Adaptive Design at KIT in April 2024. This paper has been submitted to *Computers in Human Behavior.*

Simmons, & Massey, 2015; Ku, 2020). The extensive body of literature emerging relatively quickly reveals a complex mechanism with various factors influencing the aversion or appreciation of algorithms. This demonstrates the complexity of achieving a common understanding of the underlying reasons for this behaviour.

The emergent literature consistently shows that contextual factors and personal characteristics significantly influence an individual's willingness or aversion to delegate decisions to automated systems. Building on this foundation, we explore these dimensions using a simplified multi-armed bandit problem in an experimental study. In this experiment, subjects repeatedly choose from three options with hidden expected values, aiming to identify the superior option. They can delegate their choices to a reinforcement learning algorithm at each decision point. To understand this behaviour comprehensively, we assess contextual factors by examining the effects of explainability, costs, and full task automation. Simultaneously, we analyse personal characteristics, including the Big Five personality traits, locus of control, generalised trust, and demographic information, to understand their combined impact on algorithm aversion.

Experimental studies on human-machine collaboration and algorithm aversion point to the complexity of these phenomena, influenced by a range of factors from decision consequences and task complexity to decision context framing, perceived algorithm expertise, and psychological factors like responsibility attribution. For comprehensive and interdisciplinary literature collections on algorithm aversion, systematic reviews are provided in Jussupow, Benbasat, and Heinzl (2020), Burton, Stein, and Jensen (2020), and Mahmud, Islam, Ahmed, and Smolander (2022).

Despite growing awareness of algorithm aversion, extensive research is needed to cover identified potential reasons for this behaviour. This study addresses this gap by focusing on psychological and contextual measures, as Mahmud et al. (2022) and Burton et al. (2020) suggested. Our paper provides a robust assessment of individual behaviour and personality in conjunction with variations in decision contexts. We extend the literature with the novel analysis of Big Five personality traits, locus of control, and generalised trust. These factors and demographic information form a comprehensive personality profile relevant to decision-making contexts. The Big Five traits offer insights into individual behavioural differences. Locus of control pertains to individuals' perceptions of control over their lives and decisions, influencing their algorithmic reliance. Generalised trust reflects overall trust in people and systems, affecting the acceptance of automated decisions. For contextual factors, we explore well-known elements such as information provision and task automation benefits and introduce the novel dimension of payment, assessing willingness to pay. This diverse multi-dimensional analysis offers significant insights into the literature and provides guidance for policy and future research.

Given the phenomenon's reported complexity, we began our methodological approach with statistical and regression analyses to understand treatment differences and explore variable relationships. We then used machine learning and causal inference techniques, including Logistic Regressions, Random Forests, Gradient Boosting Machines, and Uplift Random Forest classifiers, to probe the nuanced nature of decision delegation behaviour.

This paper focuses on a comprehensive analysis framework and a robust methodological construct to extract and report the effects of the analysed personality and treatment variables. Our primary objective is quantifying and reporting treatment effects in order to compare and benchmark different contextual aspects that influence algorithm aversion. Additionally, we provide insights into gender effects, particularly reactions to mistakes, and leverage the interaction with non-trained algorithms to analyse and report the learning process under different conditions. This study does not aim to assess the deeper mechanisms behind any singular of these effects but rather to analyse and compare several contextual aspects in one common framework.

Our main findings show that contextual factors like payment and automation significantly affected delegation, with payment reducing and full automation boosting its likelihood. Key personal factors influencing delegation across models included age, extraversion, openness, neuroticism, and locus of control were identified. Analysing reactions to algorithmic failures revealed that female participants were consistently more sensitive to mistakes, leading to increased distrust in algorithms after adverse outcomes. Lastly, upon analysing the learning process of the algorithm through interaction with participants, we found that higher delegation rates contributed to superior algorithm performance.

The remainder of this paper discusses related literature in section 6.2, elaborates on the experiment design in section 6.3 and the algorithm implementation in section 6.4, and provides a detailed report of all results and findings in section 6.5.

## 6.2. Related Literature

Experimental evidence on algorithm aversion and appreciation varies significantly across domains and contexts. Studies have found that different levels of human interaction with automated agents are based on factors such as task context, performance expectations, and agent roles (Chugunova & Sele, 2022). Here, we briefly discuss experimental studies and their findings regarding aversion in decision delegation to algorithms.

Studies in financial and investment contexts highlight a reluctance to fully surrender decision-making authority to automated agents despite their superior performance (Filiz, Judek, Lorenz, & Spiwoks, 2022; Gaudeul & Giannetti, 2023; Logg, 2017). Downen, Kim, and Lee (2024) found that disclosing the use of AI in financial decision-making reduces extreme investment reactions and that emotional responses, such as pleasantness and attentiveness, mediate this effect. Germann and Merkle (2023) 's experiment with young adults in financial decision-making found no significant algorithm aversion. Participants prioritised returns over the type of financial intermediary, and performance influenced their choices more than the algorithm's nature.

Human errors and significant decision outcomes seem to exacerbate algorithm aversion (Dietvorst et al., 2015; Filiz, Judek, Lorenz, & Spiwoks, 2021). Furthermore, in morally charged decisions, people often prefer the discretionary scope of human decision-makers (Jauernig, Uhl, & Walkowitz, 2022).

Heßler, Pfeiffer, and Hafenbrädl (2022) performed a contextual analysis and found that in prosocial contexts, the importance of empathy and autonomy increases algorithm aversion, leading to a preference for human-like decision support systems. Additionally, the role of responsibility sharing indicates that people may prefer human advisors because they can offload some responsibility for decision outcomes, which is less applicable to algorithmic advisors (Gazit, Arazy, & Hertz, 2023). Mahmud, Islam, and Mitra (2023) found that perceived lack of benefits (value barriers), resistance to change from established practices (tradition barriers), and negative perceptions or stereotypes (image barriers) significantly contribute to algorithm aversion among managers.

Conversely, showcasing an AI-based system's learning ability can mitigate algorithm aversion (Berger, Adam, Rühr, & Benlian, 2021). Exerting time pressure also helps reduce aversion to algorithms (Jung & Seiter, 2021). In situations where discrimination is possible, people often prefer algorithmic evaluation over human judgement (Jago & Laurin, 2022). Technology readiness also seems to attenuate these effects (Mahmud et al., 2023). For instance, Reich, Kaju, and Maglio (2023) found that a relevant driver of algorithm aversion is the misconception that algorithms cannot learn from mistakes. They demonstrated that highlighting an algorithm's learning ability significantly reduces aversion and increases trust in algorithmic predictions. These findings indicate that individuals can appreciate and even prefer algorithmic decision-making under certain conditions.

During the writing of this paper, numerous experiments with unique and elaborate designs were conducted, analysing a wide array of factors influencing algorithm aversion. Most of these studies are relatively recent, reflecting the growing interest and evolving understanding of this phenomenon. We aim to contribute to this body of knowledge by providing novel insights into psychological and contextual factors affecting algorithm aversion.

## 6.3. Experimental Design

The experimental setting employed a between-subject design, utilising a simplified version of the multi-armed bandit problem (Robbins, 1952). Our design parallels previous works, notably by Hoelzemann and Klein (2021), who also examined human interactions with bandit-based decision-making scenarios. The primary task involved participants repeatedly choosing one of three options labelled as "products" over 40 periods. The experiment was conducted online, where participants were instructed to select from three products, each with distinct hidden quality levels that represented their expected values, translated into the probability of receiving a payoff from the chosen option. The three variants of quality were low (50% chance of payoff), medium (70% chance of payoff), and high (90% chance of payoff). Participants were informed of these probability values but not which specific product they were assigned to. These probabilities were randomly assigned to products 1 to 3 at each participant's onset and remained constant throughout the experiment. Through repeated choices, the expected goal was for the participants to identify the high-quality product to maximise their total payoffs. After each selection, participants received feedback on the outcome of their decision. In each round, participants could delegate the decision to an algorithm. After reading the instructions, we asked participants about their perception of using algorithms for decision-making in regular tasks. The responses were categorised as positive, neutral, or negative. This response was used as a variable in the study, referred to as *perception*.

The earnings structure depended on performance. Participants earned 1 point for each successful round, which was converted to 0.13 euros per point. To enable immediate delegation in the payment treatment, each participant received one initial bonus point, which was also given to all other participants for fairness. Additionally, participants earned 15 points for completing the personality questionnaire.

A multi-armed bandit problem does not usually have a deterministic solution. In the experiment context, a widely applied solution concept involves following a strategy that balances exploration and exploitation (Auer, Cesa-Bianchi, & Fischer, 2002; Barto, 1997). Therefore, a rational decision-maker would explore all options to gather information about each product's probability of success. Once enough information is obtained, they would exploit the product with the highest payoff probability (90%). The reinforcement learning algorithm used in the experiment mirrors this strategy by continuously updating its probability estimates based on previous choices, thereby aligning its performance with rational decision-making principles in similar scenarios (further algorithm details are found in section 6.4).

The basic framework described above is established as the "baseline" treatment. We further introduce three treatments with different contexts — explanation, payment, and automation — to investigate the impact of explainability and transparency, willingness to pay, and complete task automation on delegation behaviour. By examining these factors, we aim to better understand user preferences and friction points in algorithmic decision-making by analysing these factors. In all treatments, we employ an attention check in a given round by displaying an animal picture below the task, which participants had to identify by the end of the task. Appendix 6.10 documents the design and the experiment screens.

### 6.3.1. Explanation Treatment

As discussed in numerous studies, transparency and explainability are key factors affecting the acceptance of algorithmic decision support. Algorithm complexity often presents these tools as "black boxes," undermining their acceptance due to the lack of understanding (De Bruyn, Viswanathan, Beh, Brock, & von Wangenheim, 2020; Enholm, Papagiannidis, Mikalef, & Krogstie, 2021; Miller, 2019; Trocin, Mikalef, Papamitsiou, & Conboy, 2021; Vlačić, Corbo, e Silva, & Dabić, 2021; Zhang, Chen, et al., 2020).

The inherent complexity in high-performing computational models poses a dilemma between accuracy and transparency, as the intricacy of these models could challenge the public's comprehension (Gilpin et al., 2018; Gunning, 2017; Herm, Heinrich, Wanner, & Janiesch, 2022). This complexity underscores practitioners' ongoing challenge in maintaining explainability (Castelluccia & Le Métayer, 2019), necessitating accessible explanations irrespective of the chosen approach. Institutions and regulators also emphasise the need for transparent algorithmic decisions (Goodman & Flaxman, 2017).

We tested the information-sharing impact on delegation in this *explanation* treatment, in which participants had access to a description of the algorithm used in the product selection task. The description aimed to be non-technical and to transmit the essence of the method behind reinforcement learning to the subjects. On the primary experiment page, the following text is displayed in a text box with a prominent design: *"Reinforcement Learning: the algorithm calculates probabilities and chooses an alternative based on the success of choices in previous rounds".* The description text remained visible during the experiment.

### 6.3.2. Payment Treatment

Exploring the less examined aspect of financial incentives in algorithm aversion, people might hesitate to pay for transparent AI if costs surpass perceived benefits (König, Wurster, & Siewert, 2022). During crises, the appeal for robo-advisors —and hence the willingness to pay— escalates due to the need for financial advice (Ben-David & Sade, 2021). Similarly, radiologists are ready to pay for AI tools that expedite diagnostics (von Wedel & Hagist, 2022).

We investigate payment's role in algorithm aversion by assigning a payment requirement to algorithmic support, termed *payment* treatment. Here, participants were informed that while they can delegate decisions to an algorithm, each delegation carries a cost of 0.10 points (one-tenth of a point), aiming to introduce the psychological aspect of payment in a way that participants easily understand. The goal was to introduce payment as a contextual variable to gauge its impact, not to explore the complexities of differential willingness to pay. The cost incurred for a decision effectively restricts algorithm support to a pay-per-use basis. The points deduction reduces the expected values of the products by the same amount, introducing a "loss" for rounds where payoffs do not materialise, as the amount is subtracted from the participant's total points.

### 6.3.3. Automation Treatment

The task complexity may persuade people to accept algorithmic decisions (Bogert, Schecter, & Watson, 2021). Bucklin, Lehmann, and Little (1998) argue that from a human standpoint, full, compared with partial, automation of decision-making processes can be very desirable in terms of efficiency, such as improving productivity and effectiveness, for better resource allocation. In essence, delegating the decision is already a form of automating, as the algorithm calculates and selects the best option based on past data. We advance this process by further automating it, reducing the overall task burden. In this way, one can analyse the subjects' behaviour toward the delegation of discrete decisions compared with the delegation of the complete task.

In the *automation* treatment, the algorithm takes over the repetitive task of product selection for 40 periods, easing the participants' effort. Unlike previous treatments requiring round-by-round delegation decisions, this feature allows continuous selection without active involvement. Participants could toggle automation on or off at any stage. If they opted for delegation, they had a 5-second window to override the decision, redirecting them to the primary selection interface. Feedback remained available after each round.

### 6.3.4. Psychological and Personal Factors

Algorithm aversion can be significantly impacted by personal factors such as personality traits, demographic features, and algorithm/task familiarity (Mahmud et al., 2022). For instance, individuals with an internal locus of control tend to resist human and AI suggestions (Sharan & Romano, 2020), and neuroticism correlates with lower trust ratings. Delegation to algorithms increases when information scarcity is present and among extroverted individuals (Goldbach, Kayar, Pitz, & Sickmann, 2019). Trust in algorithms is not static but can evolve with personal experiences (Fenneman, Sickmann, Pitz, & Sanfey, 2021), which similarly impacts attitudes toward autonomous transport (Goldbach, Sickmann, Pitz, & Zimasa, 2022).

We incorporate demographic data, the Big Five personality traits, locus of control, and trust levels into our analysis, broadening our research to encompass contextual and personal aspects of algorithm aversion. The Big Five personality traits offer a comprehensive view of human personality (Goldberg, 1990), whereas locus of control illustrates an individual's belief in their power over life events (Rotter, 1966). Generalised trust signifies an individual's confidence in the reliability and benevolence of others (Yamagishi & Yamagishi, 1994). After completing the selection task, participants proceeded to this series of personality questionnaires, including control questions (see Appendix 6.10).

## 6.4. The Algorithm: Reinforcement Learning Implementation Framework

The term “algorithm” has various definitions across fields. Computer science typically defines it as a step-by-step procedure or set of rules used to perform tasks (Cormen, Leiserson, Rivest, & Stein, 2001). In the context of algorithm aversion, it often refers to decision-making tools that assist humans in making choices or predictions (Dietvorst et al., 2015).

A variety of algorithms could be applied to the task of repeatedly selecting alternatives that maximise one's payoffs. In our design, we aimed to allow participants to observe the algorithm's training and improvement process throughout the task while keeping it simple enough for participants in the explanation treatment to understand its core mechanism in just a sentence or two. As a result, we chose the Reinforcement Learning (RL) model, a class of solution methods well-suited for learning-based and sequential problems.

Reinforcement learning is typically framed as an optimisation problem to identify optimal actions based on defined criteria (Barto, 1997). The model's framework is designed to map situations to actions in a way that maximises rewards, as defined by Sutton and Barto (2018). Critical components of reinforcement-based models include a set of choices or actions, a mechanism for receiving feedback associated with each choice, an updating rule that adjusts previous beliefs or estimates of each choice's expected value based on the feedback, and a decision rule that determines the probability of selecting each choice based on current beliefs. Our model is based on Erev and Roth (1998) 's implementation, which incorporates the concept of attractions, or weights attached to strategies that represent the perceived value associated with specific choices (C. Camerer & Hua Ho, 1999). Our implementation assigns an attraction value to each product, which is updated after a decision is made using a learning rule. The attractions are transformed into probabilities of choice using a softmax function. A formalisation of the algorithm is presented in Appendix 6.7.

The embedding of this algorithm in the experiment generates one instance of reinforcement learning for each participant, which starts with no pre-training or bias. The attraction values are initialised at 0, and the algorithm learns from participant choices and its own choices over time, making the learning process for humans and algorithms comparable.

## 6.5. Results

We conduct a comprehensive six-stage analysis of decision delegation to an algorithm, exploring its contextual, behavioural, and personal dimensions. We begin with an overview of our sample information and attention analysis, followed by examining delegation behaviour across different treatments. We then use regression methods to identify significant predictors of delegation behaviour and machine learning methods for a nuanced understanding of algorithm aversion. We incorporate causal inference methods to clarify causal relationships, analyse participants' reactions to algorithmic failures, and measure the algorithm's performance under varying conditions. This multifaceted approach provides a detailed understanding of the complex phenomenon of algorithm aversion[2]. The code and data to reproduce the experiment and all results presented hereafter are made public [3]

### 6.5.1. Sample Information and Attention Analysis

A total of 358 participants participated in our online experiment, recruited via email distribution lists sent to university students. Subjects were evenly distributed across the four treatments, with approximately 89 to 91 participants per treatment. On average, the experiment took 11 minutes to complete, and participants earned between 4 and 10 euros, with an average of 6.13 euros. Demographically, the sample was 52.7% female. Participants were primarily from Germany (51%), with the remaining individuals representing various nationalities. Most participants (73.2%) were from the Rhine-Waal University of Applied Sciences, whereas 26.8% were from Heidelberg University (both in Germany), aged between 18 and 47 years old; the mean age was 25. Among the subjects, 19% were economics students; the rest were from various other academic disciplines, of which 21% came from

[2] This research project was pre-registered in AsPredicted.org, with the ID 119401. The pre-registration covered the experimental design, treatment conditions, and standard non-parametric statistical analyses. The machine learning methods used in section 6.5.4 were decided upon after examining the data and were not included in the pre-registration.

[3] The paper's data and programming resources are available on GitHub.

STEM majors. The self-reported perception values were 46.6% positive, 43.9% neutral, and 9.5% negative.

We analysed participants' attention, mainly focusing on the automation treatment, to determine if active supervision of the algorithms' decisions persisted in a fully automated task. To measure this, we calculated the total time the web page was active in the subjects' browsers. Additionally, we implemented attention-check questions in both the experimental task and the personality questionnaires. The results are summarised in the table 6.1; these values do not account for the first round, which includes the time spent reading the instructions.

| Treatment | Average Active Time (s) | Animal Question (frequency correct) |
|---|---|---|
| Baseline | 9.6 | 0.88 |
| Explanation | 10.3 | 0.89 |
| Payment | 9.0 | 0.85 |
| Automation | 11.2 | 0.55 |

Table 6.1.: Attention metrics for all treatments

The active time analysis showed consistent results across all treatments, with participants spending an average of 9 to 11 seconds per round. A second attention check involved identifying an animal that appeared during the final rounds, revealing decreased attention in the automated treatment. Even though the screen was active, fewer people in the automated treatment seemed to monitor the task closely. We included an attention self-report question in the automated treatment, especially asking if the subject had supervised the algorithm's decisions during the task. A total of 76% answered yes, which deviates from the 55% of participants who got the animal question correct, whereas 15% answered no, and 9% answered not applicable. The difference between 76% and 55% suggests an over-reporting of the attention and supervision levels in the automated treatment. Four control questions were embedded in the personality tests, with 78% of participants answering all four correctly and 93% answering at least three correctly, indicating attentive reading.

### 6.5.2. Delegation Behaviour and Treatment Effects

We measured the frequency of delegating decisions to the algorithm in each treatment. Table 6.2 documents the absolute frequency of delegation in each treatment.

| Treatment | Frequency of Delegation |
|---|---|
| Baseline | 53.02% |
| Explanation | 58.37% |
| Payment | 27.87% |
| Automation | 66.07% |

Table 6.2.: Absolute frequencies of delegation across the four treatments

In the baseline treatment, we observed a balanced split, where about half of the decisions were delegated across participants and rounds. The information shared in the explanation treatment only slightly increased the number of delegation decisions. The introduction of payment sharply decreases, and the possibility for automation increases the willingness to allow the algorithm to decide.

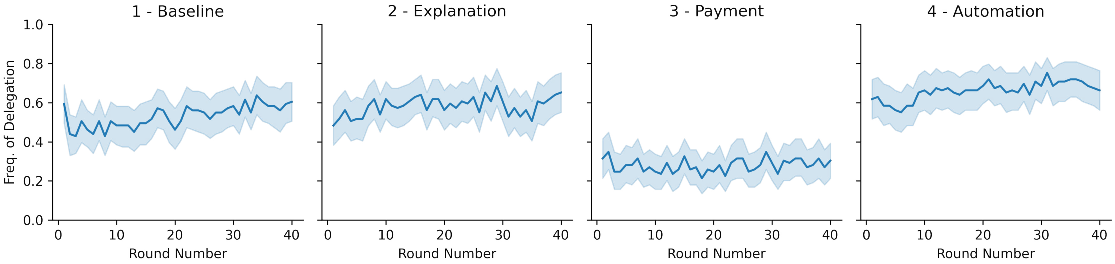


Figure 6.1.: Mean frequencies of delegation over time

Figure 6.1 displays the overall delegation frequencies over time, where the distributions are consistent across treatments and relatively constant, without any significant variations in the decision behaviour between rounds. We aggregated the experimental data on a participant level to test these findings for statistical significance. Each participant's cumulative delegation frequency over 40 periods is treated as an independent observation. The distributions of these relative frequencies of delegation are displayed in the histogram in figure 6.2.

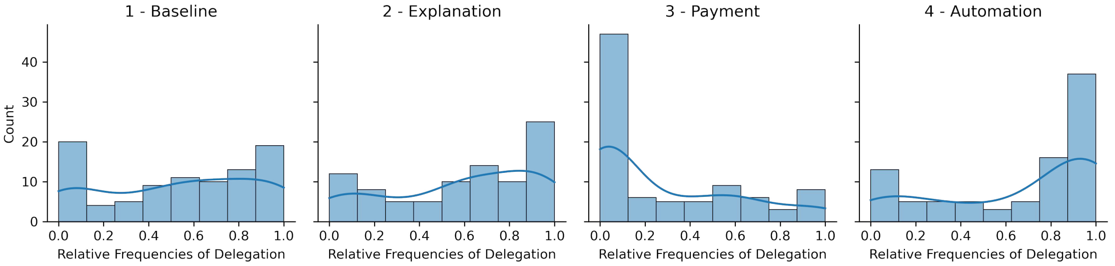


Figure 6.2.: Histogram of participants cumulative delegation frequencies

As anticipated, the highest delegation frequencies occur in automation and the lowest in payment treatments. The baseline and explanation treatments exhibit a more even distribution of subjects' delegation behaviour. We employed a Kruskal-Wallis test (Kruskal & Wallis, 1952), a non-parametric statistical test comparing the medians of several independent samples. With a test statistic of 52.67 and a p-value $< 0.001$, the results indicate a significant difference between the medians of the four independent treatment samples.

While the Kruskal-Wallis test reveals significant differences, it does not provide detailed insights into these differences between the samples. Consequently, we employed a Dunn post-hoc test (Dunn, 1961) to identify significant pairwise differences between samples. The p-values for these comparisons are in table 6.3.

In summary, these results suggest significant differences between the medians of baseline, payment, and automation, as well as between explanation and payment. There is no significant difference between the medians of baseline and explanation or between explanation

| | Baseline | Explanation | Payment | Automation |
|---|---|---|---|---|
| Baseline | 1 | | | |
| Explanation | 0.373 | 1 | | |
| Payment | $< 0.001$ | $< 0.001$ | 1 | |
| Automation | 0.009 | 0.090 | $< 0.001$ | 1 |

Table 6.3.: Dunn posthoc test results, p-values for pairwise treatment comparisons

and automation. The payment feature was the most influential regarding the willingness to delegate.

The contextual findings highlight the influence of different treatment conditions on the delegation behaviour of participants. The baseline and explanation treatments led to a more even distribution of delegation behaviour. On the other hand, the payment treatment had a considerable negative impact on the willingness to delegate. The automation treatment led to the highest frequency of delegation among the four treatments, demonstrating the importance of reducing the workload involved in a task to encourage algorithm-based decision-making. Overall, these results underscore the significance of understanding and addressing the factors that affect delegation behaviour to design more effective human-algorithm collaborations and decision-making processes.

### 6.5.3. Incorporating the Personal Variables - Regression Analysis

The design of our treatments provides insights into how exogenous factors influence delegation behaviour. However, individual factors also play a significant role in algorithm aversion, as widely discussed in the literature. In this section, we examine the binary action of delegating a decision in relation to treatment conditions and personal factors, including personality test scores, gender, education, and self-reported perception (as explained in section 6.3). Categorical values were encoded as binary dummy variables.

Although correlations between the variables under investigation and delegation are primarily weak, they are highly significant (full correlation results are reported in 6.9, Appendix 6.8). We constructed a logistic regression (LR) model to explore further and quantify these relationships, including demographic and personal information as independent variables. The model results are summarised in table 6.4. A critical remark in the regression modelling is that we use the entire experiment's dataset: every decision from each participant at each round. Due to repeated choices made by the same individuals across 40 periods, we clustered the standard errors on the participant level. This approach accounts for intra-participant correlation, considering potential influences from unobserved individual factors or shared experiences, as per Bertrand, Duflo, and Mullainathan (2004)'s reasoning.

The logistic regression model provides several insights into the effects of treatments and personality traits on delegation and also reinforces the findings in section 6.5.2. Initially, the automation treatment exhibits a positive and statistically significant impact on delegation ($p = 0.027$), suggesting that automating tasks encourages individuals to delegate. Conversely, the payment treatment displays a negative and statistically significant influence ($p < 0.001$), implying that requiring payment could discourage delegation. The

| Variable | Coefficient | Standard Error | p-value |
|---|---|---|---|
| Constant | $-0.525$ | 1.013 | 0.605 |
| Explanation | 0.252 | 0.207 | 0.223 |
| Payment | $-1.012^{***}$ | 0.235 | $< 0.001$ |
| Automation | $0.515^{*}$ | 0.234 | 0.027 |
| Female | $-0.144$ | 0.179 | 0.421 |
| Age | $-0.009$ | 0.018 | 0.596 |
| STEM | 0.267 | 0.227 | 0.238 |
| Business & Economics | $-0.181$ | 0.201 | 0.37 |
| Extraversion | 0.04 | 0.059 | 0.497 |
| Agreeableness | 0.036 | 0.073 | 0.627 |
| Conscientiousness | 0.137 | 0.085 | 0.106 |
| Neuroticism | $-0.047$ | 0.072 | 0.513 |
| Openness | $-0.008$ | 0.087 | 0.926 |
| Internal LoC | 0.057 | 0.102 | 0.578 |
| External LoC | 0.054 | 0.106 | 0.614 |
| Generalised Trust | 0.067 | 0.066 | 0.307 |
| Perception | $-0.368^{**}$ | 0.14 | 0.009 |

Table 6.4.: Logistic regression results - delegation

$R^2$=0.08. Significance levels: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

explanation treatment, although positive, is not statistically significant ($p = 0.223$). Regarding personal variables, the only statistically significant effects are observed for perception ($p = 0.009$)[4], which negatively impacts delegation, suggesting that an increase in negative perception about algorithms is correlated with a lower likelihood of delegation. Other variables, including gender, age, field of study, and personality traits, do not exhibit statistically significant effects on delegation in this model. A second regression model, including interaction terms, is reported in Appendix 6.8, in which payment loses significance, and the internal locus of control becomes significant. Quantile regression models applied to cumulative delegation frequencies (shown in figure 6.2) showed similar significance and coefficients to logistic regression despite a marginally better fit. See Appendix 6.8 for complete details.

In conclusion, examining personality traits and algorithm aversion uncovers the influence of individual factors and treatment conditions on delegation behaviour. A critical insight from this analysis is the existence of intricate relationships between various traits. Interaction terms offer a more comprehensive understanding of the relationships between variables and delegation behaviour by accounting for the dependence of some variables' effects on the values of other variables. Gaining insights into these relationships can aid in comprehending how diverse behavioural profiles respond to algorithmic systems.

In conclusion, the regression analyses reveal the influences of individual factors and treatment conditions on delegation behaviour, particularly regarding automation, payment, and perception. While most individual characteristics, such as gender, age, and specific

[4] As a reminder, perception refers to the participants' self-reported views on delegating everyday decisions to algorithms, ranging from negative to neutral to positive.

personality traits, did not show significant effects on their own, the interaction terms in our extended model (see Appendix 6.8) suggest that the impact of some variables depends on the presence of others. Despite these findings, the overall model explains only approximately 8% of the variance (Pseudo-$R^2 = 0.08$), indicating the complexity of the relationship between variables. Furthermore, the relatively low predictive performance of the logistic regression models (around 69% ROC-AUC, details in the next section) suggests that simple models do not fully capture these relationships. This complexity and the need for a more robust analysis led us to adopt ensemble learning models to better account for intricate variable relationships.

### 6.5.4. Machine Learning for Delegation Behaviour Analysis and Causal Inference

To understand whether personal and contextual information helps predict delegation behaviour in such a case, we tested a few prediction techniques using the same variables scheme, that is, predicting the binary outcome of the delegation decision possibility using treatments, personality, and demographic data.

C. F. Camerer (2018) highlights the benefits of applying machine learning to model behaviour, emphasising its potential for improved predictive accuracy, handling large datasets, capturing non-linear relationships, and adaptability. Additionally, machine learning enables personalisation and fosters cross-disciplinary insights, contributing to a better understanding of human decision-making and facilitating more effective interventions across various domains.

The logistic regression model, as detailed in section 6.5.3, offers limited insights into the complex interplay of our variables, accounting for only approximately 9% (pseudo R-squared) of the variation in delegation decisions. Given the absence of clear linear relationships and the complexity of the data, we turn to more sophisticated methods. We employ machine-learning models to examine the overall impact of variables on predicting delegation, followed by causal machine learning models to separate treatment effects from the personal covariates. In the subsequent models, we refer to within-sample predictions, using 80% of the sample for model training and the other 20% to generate and test predictions. Methodological formalisations for the adopted methods can be found in Appendix 6.7, and technical model implementation remarks in Appendix 6.9.

#### 6.5.4.1. Predicting Delegation Behaviour

If we use our logistic regression coefficients to generate predictions, the model yields an accuracy score of 0.62, meaning 62% of the delegation decisions were classified correctly, not far from a random baseline. This relatively low accuracy might be due to several factors influencing the results that have yet to be accounted for or the failure of the model to capture complex relationships between the variables. To deepen the understanding of these variables' relationships and the possibility of generating predictions for algorithm aversion behaviour using contextual and personal information, we resort to the machine learning techniques Random Forest and Gradient Boosting Machines.

Research shows successful predictions of behavioural elements using personality traits, characteristics, and environmental data. Balakrishnan, Khan, Fernandez, and Arabnia (2019) used psychometric test data, including Big 5 and Dark Triad, and Twitter features to predict cyberbullying accurately. Guntuku, Yaden, Kern, Ungar, and Eichstaedt (2017) employed machine learning to predict mental health status based on social media and personality data. Similarly, Stachl et al. (2017) used personality traits to predict smartphone usage behaviour. These studies demonstrate the potential of machine learning models in similar prediction tasks.

Breiman (2001) introduced the Random Forest model, an ensemble learning method for classification and regression problems. The algorithm creates multiple decision trees, each of which 'votes' on an answer. In a classification problem like ours, the Random Forest chooses the class that gets the most votes from all the trees. The key idea behind Random Forest is to create a "forest" of diverse decision trees constructed from random subsets of training data and features. This approach helps increase the model's robustness, reduce overfitting, and improve overall predictive accuracy. The Random Forest algorithm is particularly useful for binary classification problems because it can handle non-linear relationships between the input features and the output variable. It can also handle missing values and outliers in the input data and estimate the importance of each input feature in the prediction (Liaw, Wiener, et al., 2002).

Similarly, Gradient Boosting Machines (GBMs) are a class of ensemble learning algorithms that build a robust model by iteratively adding weak learners, typically decision trees, to minimise a loss function. The algorithm focuses on correcting the errors of the previous tree by training on the residuals, effectively improving the overall model's performance, as defined in Friedman (2001).

As per definitions by Breiman (2001) and Friedman (2001), Random Forest and GBMs are ensemble learning methods for similar purposes. The main difference lies in their approach to building the ensemble of decision trees. Random Forest constructs multiple trees independently and in parallel, combining their predictions through averaging or majority voting. It uses bagging (Bootstrap Aggregating) to create diverse trees by resampling the dataset with replacement. In contrast, GBM constructs trees sequentially, with each new tree trying to correct the errors made by the previous tree. It utilises a technique called boosting, where trees are combined through a weighted majority vote, and the weights are determined by minimising a loss function during the training process. We apply both methods for comparable results but with distinct processes, enabling comparison and validation of the findings from the generated predictions to assess our findings' consistency. In each model, feature importances highlight the significance of each feature in predicting the target variable. Figure 6.3 presents an overview of the feature importances for both models aggregated by the mean (for separate plots for each model, see Appendix 6.8.3).

Both models have been cross-validated during parameter fitting and training to avoid overfitting (details in Appendix 6.9). Our cross-validation procedure grouped observations on the participant level, ensuring instances of the same participant in either the training set or the test set.

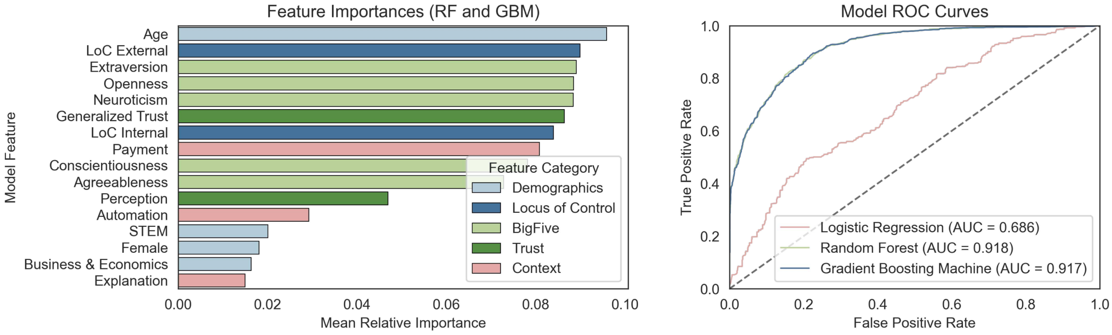


Figure 6.3.: Machine learning model results: Feature importances and ROC curves

Upon analysing the model feature importances, it becomes clear that no single or small group of features dominates the influence on decision delegation. This influence unfolds into a mix of variables that include both personal and contextual elements. Since feature importances only outline effect size without indicating direction, we perform the following analysis alongside the regression coefficients.

Age emerges as the most significant variable, reflecting its essential role in shaping comfort with algorithmic delegation, which aligns with the findings from Germann and Merkle (2023). The regression analysis indicates that older individuals are less likely to delegate to algorithms. External locus of control is the second most important factor, suggesting that individuals who believe external forces determine their outcomes are more likely to delegate decisions to algorithms. The positive coefficient in the regression models supports this.

Extraversion, another prominent factor, suggests that more outgoing and social individuals may be more comfortable delegating tasks to algorithms, consistent with a positive influence. Openness, however, shows mixed effects. While it positively influences correlation, it appears negative in some regression models, indicating a nuanced relationship. Neuroticism also shows mixed effects, with a generally positive influence in some models but negative in correlation, reflecting the complexity of its impact on delegation behaviour.

Generalised trust reflects overall trust in people and systems and significantly impacts the likelihood of delegating decisions to algorithms. High trust correlates with a higher propensity to delegate, as indicated by the positive coefficient in the regression analysis. While important, the internal locus of control has a smaller impact than the external locus of control, showing that individuals who feel more in control of their outcomes are less likely to delegate to algorithms.

Contextual factors like payment and automation also play significant roles. Payment decreases the likelihood of delegation, suggesting that individuals are less willing to delegate decisions when a cost is involved, as evidenced by the solid negative coefficient. Full task automation increases the likelihood of delegation, underscoring a preference for fully automated systems in decision-making processes, supported by a positive coefficient.

We evaluated the Logistic Regression (LR), Random Forest (RF), and Gradient Boosting

Machine (GBM) models using four metrics: Accuracy, Precision, Recall, and F1 score. Accuracy calculates the proportion of correctly classified instances. Precision quantifies how well the model correctly identifies positive instances. Recall gauges the model's ability to detect positive instances among actual positives. The F1 score, a blend of precision and recall, is the harmonic mean of these two metrics (Powers, 2020; Sokolova & Lapalme, 2009). As summarised in Table 6.5, both RF and GBM outperformed LR in predictive power, with RF achieving slightly superior performance across all metrics. This outcome highlights the efficacy of tree-based models for our classification problem.

| | LR | RF | GBM |
|---|---|---|---|
| Accuracy | 0.6210 | 0.8332 | 0.8325 |
| Precision | 0.6112 | 0.8185 | 0.8120 |
| Recall | 0.7018 | 0.8639 | 0.8639 |
| F1-score | 0.6534 | 0.8406 | 0.8414 |

Table 6.5.: Prediction performance metrics

In addition, a Receiver Operating Characteristic (ROC) curve provides a graphical representation of a classifier's performance across varying decision thresholds (figure 6.3, right-hand side). The Area Under the ROC Curve (AUC-ROC) measures the overall performance of a binary classifier. It ranges from 0 to 1, with higher values indicating better performance. 0.5 indicates a random classifier (dashed line), and 1 indicates a perfect classifier. The ROC area quantifies how well the classifier can distinguish between the positive and negative classes, regardless of the choice of classification threshold (Bradley, 1997; Fawcett, 2006). In the overall analysis, and in line with previous performance metrics, the LR model is surpassed by the other models, with the RF model showing a slight edge. The high scores achieved by both the RF and GBM models are due to their ability to explain the data, enhancing the reliability of the interpretations documented in our study.

Although logistic regression provided valuable insights into the direction and significance of individual variables, its ability to handle the complex data relationships in our study was limited. We explored machine learning techniques to better capture these relationships, specifically Random Forest and Gradient Boosting Machines. Both models significantly outperformed logistic regression in terms of accuracy, precision, recall, and F1 score, with the Random Forest model having a slight edge in accuracy over the GBM. Both models consistently highlighted the same features, such as payment, extraversion, and neuroticism, as key influencers in delegation decisions.

#### 6.5.4.2. Causal Inference and Heterogeneous Treatment Effects - Uplift Random Forest

To further understand the factors influencing decision delegation to algorithms, we now focus on disentangling the effects of the treatment conditions from personal data. While regression and machine learning models have provided insights, they combine all variables, not distinguishing between treatment conditions and personal characteristics effects. Hence, we use causal inference to uncover how treatment effects vary across different subgroups within our sample, focusing on estimating the expected change in the outcome as a result of the intervention. This approach allows us to measure heterogeneous treatment effects and identify the subset of individuals most influenced by the treatment conditions, given their characteristics. To this end, we resort to Uplift Modelling.

Uplift Modelling, a branch of causal inference, models the impact of incremental treatment effects on individuals' behaviour (N. J. Radcliffe & Surry, 2011). Early applications of similar methods can be seen in N. Radcliffe and Surry (1999). For a comprehensive definition and literature review on machine learning problems and applications, see Gutierrez and Gérardy (2017); N. J. Radcliffe and Surry (2011).

We employ the Uplift Random Forest Algorithm, an ensemble learning method that uses the random forest algorithm to estimate the causal effect of a treatment or intervention on individual outcomes (Guelman, Guillén, & Pérez-Marín, 2012, 2015). The uplift random forest classifier (Sołtys, Jaroszewicz, & Rzepakowski, 2015) incorporates the treatment indicator as a covariate to capture differential effects and uses other covariates to estimate individual treatment effects. The model is tuned using the same cross-validation technique described in 6.5.4.1, with details in Appendix 6.9.

Treatment effects can be evaluated at an individual level by computing uplift scores. These scores represent the predicted likelihood of delegation for each observation under each treatment scenario, essentially providing a probabilistic estimate of how a participant would behave if they were subjected to a specific treatment. The distributions of these predicted likelihoods are plotted in figure 6.4. The trend observed in this analysis follows the initial assessment of the treatment effects (section 6.5.2) in reference to the baseline. Payment negatively impacts the likelihood of delegation, whereas explanation has a slight positive effect, and automation has a more pronounced positive effect. Each treatment's computed average treatment effects are payment = −0.24, explanation = 0.07, and automation = 0.15.

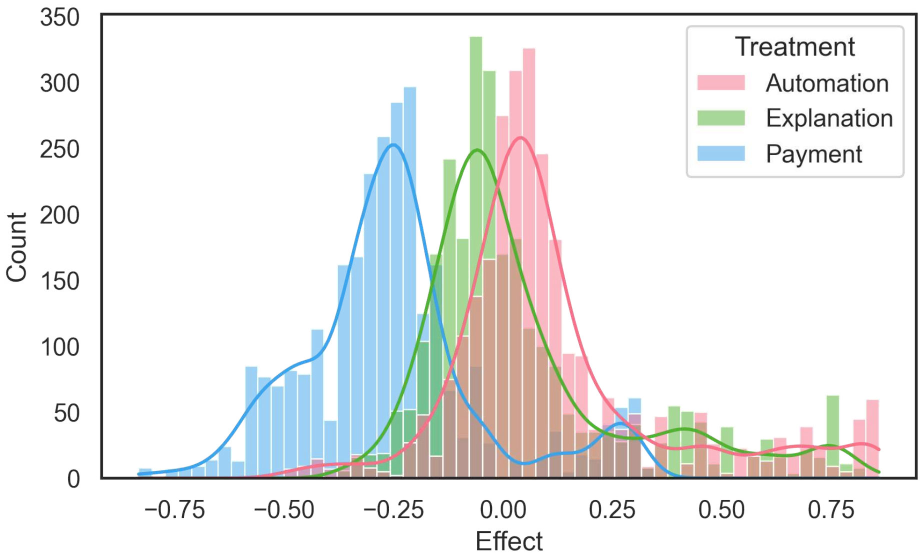


Figure 6.4.: Distribution of predicted treatment effects (Uplifts)

Feature importance can also be extracted from this model, with a slightly different meaning. Unlike traditional classification models, in Uplift models, feature importance does not directly equate to the effect of a feature on the outcome but rather its influence on the treatment effects. In other words, an essential feature in the model translates to the influence on the change in the likelihood of delegation mediated by the treatment. These values are presented in figure 6.5.

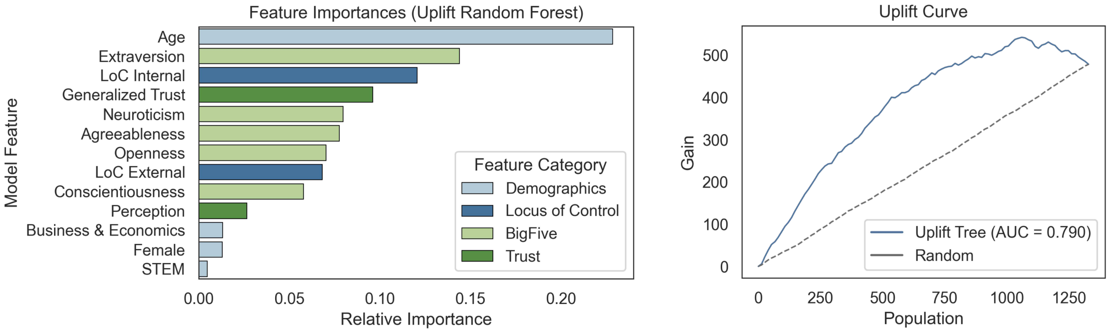


Figure 6.5.: Uplift random forest results: Feature importances and uplift curve

The Uplift Random Forest feature importances quantification aligns with the previous models, with age, extraversion, internal locus of control, generalised trust, and neuroticism significantly influencing the mediation of treatment effects on delegation behaviour. Age emerges as the most significant factor again, indicating its critical role in how treatments are received. Extraversion and generalised trust suggest that more outgoing individuals and those with higher trust in people and systems are more responsive to treatments.

Internal locus of control and neuroticism also play crucial roles, with individuals who feel more in control of their outcomes and those with higher emotional instability responding differently to treatments. Contrasting with the Random Forest and Gradient Boosting Machine models, the Uplift model highlights these factors' roles in optimising treatment effects rather than the outcome itself. These findings emphasise the nuanced interplay of personal traits and contextual influences in mediating delegation decisions to algorithms.

Unlike traditional machine learning models, which compare predicted outcomes to observed labels, causal models predict the difference between observed outcomes and unobserved counterfactuals. This makes using standard classification metrics like precision, recall, or ROC AUC (as in 6.5) impossible. Instead, we use metrics specific to uplift models, such as uplift curves, which plot the cumulative gain from targeting individuals based on predicted uplift. The Area Under the Uplift Curve (AUUC), analogous to the AUC-ROC, measures the model's ability to prioritise effective interventions. Figure 6.5 (right-hand side) shows our model's Uplift Curve.

We have computed the AUUC using a synthetic control group consisting of individuals whose predicted optimal treatment matches the actual treatment they received or those in the actual control group, following the method by H. Chen, Harinen, Lee, Yung, and Zhao (2020). The uplift score for each individual in the synthetic control was computed, and individuals were ranked based on these scores. The AUUC was then calculated as the area under the curve plotting the cumulative proportion of actual outcomes against the proportion of the population targeted. The result is 0.79, indicating relatively high performance in the prediction task and explanation power.

Applying Uplift Random Forest show that most variables affecting delegation behaviour also mediate the impact of treatments on these decisions, strengthening the influence of identified variables such as age, extraversion, internal locus of control, generalised trust, and neuroticism in the observed behaviour.

### 6.5.5. How Subjects React to Non-Profitable Algorithmic Decisions

Numerous studies show that people initially trust algorithms, but trust may plummet after a mistake occurs (Glikson & Woolley, 2020). Dietvorst et al. (2015) found that people avoid algorithms or computerised decision-making systems even if they make fewer errors than humans due to high expectations for algorithms and attributing errors solely to the algorithm. Prahl and Van Swol (2017) showed that people are less likely to follow advice from a computer algorithm immediately after receiving incorrect advice. In addition, Chong, Zhang, Goucher-Lambert, Kotovsky, and Cagan (2022) reveals that poor algorithmic performance harms human confidence in the algorithm and self-confidence. Bogert et al. (2021) agree with the idea of adverse reactions by outlining that bad decisions generated by algorithms are more severely punished than those of humans. To investigate this further, we analysed participants' reactions after delegating a decision to the algorithm and receiving no payoff.

Regarding the impact of the algorithms' performance on the subjects, we calculated the frequency of participants changing their strategies from "delegate" to "not delegate" relative to the number of times the algorithm's decision resulted in a zero payoff, which does not necessarily mean a "wrong" choice but can also indicate a non-realised payoff from the "correct" choice. We extended this analysis to explore potential gender effects. Table 6.6 presents the absolute proportions of reaction results categorised by gender and treatment.

| | Baseline | Explanation | Payment | Automation |
|---|---|---|---|---|
| General (aggregated) | 0.30 | 0.25 | 0.35 | 0.09 |
| Male | 0.26 | 0.15 | 0.27 | 0.07 |
| Female | 0.34 | 0.31 | 0.40 | 0.10 |

Table 6.6.: Relative frequencies of changing strategies (reaction) following algorithmic failures

On average, participants in the payment treatment group exhibited the highest reaction frequency (0.35), suggesting that individuals are more likely to change their decision when a financial incentive is involved. Conversely, the automation treatment group had the lowest frequency of reaction (0.09), indicating that participants are less likely to change their decision when the task is automated, possibly due to the complete handover process or also satisfaction with the algorithm performance, which was overall higher in the automation treatment (further details on the algorithm's performance are documented in section 6.5.6).

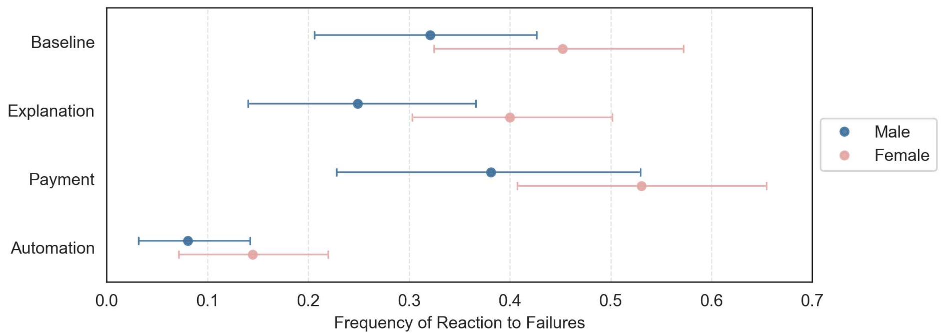


Figure 6.6.: Frequencies of reaction to algorithmic failures by treatment and gender

Comparing reaction frequencies between males and females reveals that females have a higher reaction frequency across all treatments, suggesting they might be more sensitive to algorithm mistakes (figure 6.6). Given that gender differences were not observed elsewhere in the experiment, we conducted statistical tests on both samples to further examine the gender gap in reaction.

Similar to the statistical tests performed on the relative frequencies of delegation, we calculated the relative frequencies of reaction for each participant over 40 periods, treating each participant's decision path as an independent observation and separating the samples by gender. We then applied a Mann-Whitney U test (Mann & Whitney, 1947) to measure the difference between the two independent samples. The results show a value of 7751.51 and a p-value of 0.0028, outlining a statistically significant difference between the means of the frequency of strategy reactions for males and females. To deepen our understanding of participant reactions, we further analysed whether contextual or personal factors influenced their behaviour. Similar to the methodology used in the delegation behaviour analysis (section 6.5.3), we employed a logistic regression with standard errors clustered at the participant level. The results of this analysis are compiled in Table 6.7.

The analysis indicates that task automation, gender, and internal locus of control are key factors in strategy changes following unprofitable algorithm decisions. Full task automation and a high internal locus of control reduce the likelihood of strategy shifts, suggesting trust in the process and personal control beliefs. Conversely, female participants are more prone to strategy changes, hinting at potential gender differences in reactions to algorithmic failures. Other factors, including algorithm explanation, payment requirement, and various personality traits, do not significantly influence strategy changes, suggesting their impact may be less direct.

### 6.5.6. Task Performance and Human-Algorithm Interaction

To evaluate how well the reinforcement learning (RL) algorithm performed in the product selection task, we analysed the evolution of the mean probabilities assigned to selecting each product quality level (low, medium, high) over time. These probabilities were updated each round based on the outcomes, and we plotted these changes to visualise the

| Variable | Coefficient | Standard Error | p-value |
|---|---|---|---|
| Constant | $-2.188^{*}$ | 0.868 | 0.012 |
| Explanation | $-0.088$ | 0.181 | 0.626 |
| Payment | $-0.32$ | 0.214 | 0.134 |
| Automation | $-1.042^{***}$ | 0.24 | $< 0.001$ |
| Female | $0.453^{**}$ | 0.154 | 0.003 |
| Age | 0.005 | 0.016 | 0.775 |
| STEM | $-0.345$ | 0.213 | 0.106 |
| Business & Economics | 0.021 | 0.182 | 0.908 |
| Extraversion | 0.031 | 0.053 | 0.567 |
| Agreeableness | 0.121 | 0.081 | 0.138 |
| Conscientiousness | 0.027 | 0.082 | 0.743 |
| Neuroticism | $-0.044$ | 0.061 | 0.465 |
| Openness | 0.04 | 0.09 | 0.657 |
| Internal LoC | $-0.292^{***}$ | 0.087 | $< 0.001$ |
| External LoC | $-0.107$ | 0.104 | 0.304 |
| Generalised Trust | $-0.034$ | 0.07 | 0.63 |
| Perception | $-0.229$ | 0.122 | 0.06 |

Table 6.7.: Logistic regression results - reactions

Pseudo-$R^2 = 0.03$. Significance Levels: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

algorithm's learning process. The experimental task presented a challenge due to the possibility of receiving a zero payoff even after identifying the best option, which could alter the perceived value of the correct choices. This added ambiguity influenced both the human participants and the algorithm's ability to converge on the optimal solution.

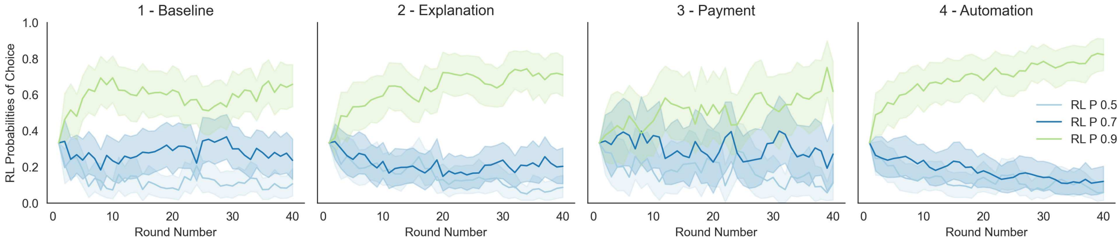


Figure 6.7.: Reinforcement Learning choice probabilities over time

In summary, the participants' degree of trust and delegation influenced the RL algorithm's learning process and success in identifying the optimal product. Higher trust in the algorithm, as seen in specific treatments, corresponded to better performance in selecting the highest-quality product.

Table 6.8 compares the algorithm's and human subjects' performance throughout the task. The values denote the success frequencies, normalised by the number of human or algorithm decisions. Even without prior training and learning on the spot, the algorithm outperformed the human subjects in most cases.

Figure 6.7 shows how the choice probabilities for each product type developed over the experiment. The RL algorithm could generally identify the highest-quality product more

| | Algorithm | Human |
|---|---|---|
| Baseline | 0.592 | 0.511 |
| Explanation | 0.634 | 0.495 |
| Payment | 0.506 | 0.505 |
| Automation | 0.694 | 0.515 |

Table 6.8.: Frequency “high” product selected

effectively than the inferior options. However, the algorithm’s performance varied across different treatment conditions. We observed that the algorithm’s ability to generate optimal choice probabilities improved in treatments with higher delegation rates, such as the explanation and automation conditions. Conversely, the payment treatment negatively impacted performance, likely due to the costs associated with using the algorithm.

## 6.6. Conclusion and Discussion

This study examined how contextual variables and personal characteristics influence delegation behaviour in algorithmic decision-making. By exploring the effects of explainability, economic costs, and automation, alongside the impact of personality traits such as the Big Five, locus of control, and generalised trust, we identified and quantified the primary drivers of algorithm aversion. Our approach introduces novel insights by integrating payment as a contextual factor and providing a comprehensive analysis of personality traits, directly addressing existing gaps in the literature, identified by Mahmud et al. (2022) and Burton et al. (2020).

The main objective of this study was to robustly identify and quantify these effects using a multi-method analysis that includes hypothesis testing, regression, machine learning, and causal inference models. We found that:

- **Contextual factors:** Clear explanations and complete automation significantly increased delegation, enhancing trust and reducing the decision-making load. In contrast, introducing a fee reduced delegation, highlighting cost sensitivity.
- **Personal characteristics:** Age, locus of control, extraversion, openness, and neuroticism significantly and consistently influenced delegation decisions. These traits contributed in a balanced manner, without any single factor predominating, highlighting the equitably distributed influence among personality and contextual factors.
- **Mediation effects** Age, extraversion, locus of control, trust, and neuroticism mediate the impact of contextual treatments.

The application of both traditional machine learning models and Uplift Random Forest highlighted not only the significant predictors of delegation behaviour but also the nuanced effects of contextual adjustments on different subgroups. These findings are further supported by high predictive success metrics in cross-validated settings, which demonstrate the models’ effectiveness and accuracy in reflecting complex behavioural dynamics.

Our study aligns with prior research in several points. This interaction of personal and contextual elements is consistent with findings by Snijders, Conijn, de Fouw, and van Berlo (2023). The negative response to payment observed in our study reflects Mahmud et al. (2023)'s findings on value and tradition barriers contributing to aversion, while the increased delegation under full automation parallels Jung and Seiter (2021)'s observation that reduced decision-making burdens can lessen aversion. Age significantly influenced delegation decisions, supporting Germann and Merkle (2023)'s demographic insights. Additionally, our observations on gender and emotional traits like neuroticism resonate with Downen et al. (2024)'s findings on emotional responses mediating AI interactions.

Practically, our results underscore the importance of designing algorithmic systems adaptable to user-specific characteristics and contextual details. Such systems should cater to the variability in user preferences and economic sensitivities, thus enhancing the effectiveness of human-algorithm collaboration. Moreover, our findings highlight that negative reactions to algorithmic errors, particularly in payment scenarios and predominantly among females, provide crucial insights for developing strategies to maintain user trust and engagement. For instance, higher trust levels in treatments such as explanation and automation were associated with better performance by the RL algorithm in selecting high-quality options.

However, the study's limitations, including the simplicity of the experimental design and potential non-representativeness of the sample, suggest caution in generalising these findings. Future research can explore more complex decision-making tasks within realistic settings and employ a broader range of algorithmic models, varying in training settings and parameters. Additionally, it would be beneficial to experiment with variations in our treatments, such as adjusting payment levels and providing explanations with varying levels of detail. Investigating new variables and contextual factors could also provide deeper insights into the dynamics of human-algorithm interaction, further enhancing our understanding of these systems.

In conclusion, this paper advances our understanding of the critical factors that shape algorithm aversion and decision delegation. By highlighting the roles of contextual framing and individual differences, it lays the groundwork for future studies to explore the broader policy implications and practical applications of these findings. We encourage further investigation into how specific characteristics influence decision-making behaviours in various settings, aiming to develop decision support systems that are both user-centric and effective in diverse domains.

# Appendix

## 6.7. Methodological Formalisations

This section provides an overview of the machine learning methods used in the project. The following sub-sections account for the Random Forest, Gradient Boosting, and Uplift Random Forest methods, providing generalisations of the algorithms' implementations.

### 6.7.1. Reinforcement Learning Implementation and Tuning

The underlying problem introduces three options or products, expressed as $Q_i$, each associated with distinct probabilities of receiving a payoff that can be selected at each period, $t$. Each product $Q_i$ is associated with an attraction value $A_{Q_i}(t)$, representing the decision weight attached to product $Q_i$ at period $t$. Following the theoretical frameworks in C. Camerer and Hua Ho (1999); Erev and Roth (1998), the attraction values are updated based on the payoffs received by selecting product $Q_i$ using the following update rule:

$$A_{Q_i}(t) = \phi A_{Q_i}(t-1) + I(Q(t) = Q_i)\pi_{Q_i}(t) \tag{6.1}$$

This model features the indicator function, which means that a player's attraction to a strategy can only increase if they choose it. The attraction increases by the amount of payoff received from it. In the update rule, the indicator functions $I(Q(t) = Q_i)$ equals 1 if a participant chooses product $Q_i$ at period $t$ and 0 otherwise, while $\pi_{Q_i}(t)$ represents the payoff received when choosing product $Q_i$ at period $t$. The recency parameter $\phi$ indicates how quickly past payoffs are forgotten, which acts as a form of learning rate. Attractions from the previous period determine choice probabilities in any period. A logistic transformation over the attraction values calculates the probabilities:

$$P_{Q_i}(t+1) = \frac{e^{\lambda A_{Q_i}(t)}}{\sum_{k=1}^{m} e^{\lambda A_{Q_k}(t)}} \tag{6.2}$$

In this equation, $P_{Q_i}(t+1)$ represents the probability of selecting product $Q_i$ at time $t+1$, $A_{Q_i}(t)$ denotes the attraction of product $Q_i$ at time $t$, and $m$ indicates the number of available product options. The second parameter, $\lambda$, reflects the sensitivity of choice probabilities to differences in attractions. The two necessary parameters were tuned using observed data from 1000 simulations, testing for the ranges $0-1$ for $\phi$ and $0-10$ for $\lambda$. The tuning resulted in $\phi = 0.47$ and $\lambda = 4.5$, associated with higher payoffs. The experiment parameters were set to these values statically.

## 6.7.2. Random Forest

The Random Forest algorithm concept builds a large collection of de-correlated decision trees and then aggregates them through a majority voting system for classification problems. Hastie, Tibshirani, Friedman, and Friedman (2009) generalised the algorithm as follows:

**Algorithm 1** Random Forest Algorithm

```
Require: $B$ trees to be grown, $N$ size of bootstrap sample, $M$ total variables, $m$ selected variables, $n_{\min}$ minimum node size
Ensure: Output the ensemble of trees $\{T_b\}_1^B$
for $b = 1$ to $B$ do
    Draw a bootstrap sample of size $N$ from the training data
    Grow a decision tree $T_b$ on this data by:
    while each terminal node of the tree until the minimum node size $n_{\min}$ is reached do
        Select $m$ variables at random from all $M$ variables
        Pick the best variable/split-point among the $m$
        Split the node into two daughter nodes
    end while
end for
To make a prediction for a new point $x$, let $\hat{C}_b(x)$ be the class prediction of the $b$th random forest tree
The random forest chooses $\hat{C}_{\text{rf}}(x) = \text{majority vote}\{\hat{C}_b(x)\}_1^B$
```

More details on the Random Forest algorithm can be found in Breiman (2001).

## 6.7.3. Gradient Boosting Machines

Gradient Boosting Machines (GBM) is a machine learning method that builds a sequence of decision trees, each correcting its predecessor's mistakes, to create a final, robust predictive model (Friedman, 2001). Hastie et al. (2009) also provides a generalisation of this model, with the step-wise algorithm defined as:

**Algorithm 2** Gradient Boosting Machines Algorithm (Generalised)

```
Require: $M$ iterations, $n$ number of observations, $L$ loss function, $y_i$ observed response, $F(x_i)$ predicted response, $h_m(x)$ base learner
    at iteration $m$
Ensure: Output $F_M(x)$ as the final model
1: Initialise the model with a constant value:
       $F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma)$
2: for $m = 1$ to $M$ do
3:     Compute pseudo-residuals:
           $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}, \quad \text{for } i = 1, \ldots, N.$
4:     Fit a base learner $h_m(x)$ to pseudo-residuals, i.e., train it using the training set $\{(x_i, r_{im})\}_{i=1}^n$
5:     Compute multiplier:
           $\gamma_j m = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i))$
6:     Update the model:
           $F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$
7: end for
```

Lines 2-6 are repeated $K$ times at each iteration $m$, once for each class. For a more detailed description of the Gradient Boosting Machines and their derivations, see the comprehensive overview in Hastie et al. (2009).

### 6.7.4. Uplift Modelling

The underlying method is the same as that of the Random Forest. However, For the uplift random forest classifier, the uplift tree consists of a combination of methods based on uplift modelling, with the tree split criterion based on differences in the uplift. In the standard notation (Rubin, 1974), we consider $Y_i(1)$ an individual's $i$ being treated and $Y_i(0)$ for being in the control group. In this case, the causal effect $\tau_i$ is given by $\tau_i = Y_i(1) - Y_i(0)$. Having $W_i \in 0, 1$ as a binary variable indicating if person $i$ is in the active treatment group, and 0 otherwise (control group), the observed outcome is $Y_i^{obs} = W_i Y_i(1) + (1 - W_i)Y_i(0)$.

Based on Gutierrez and Gérardy (2017), considering a balanced randomised experiment, the average treatment effects (uplifts) are estimated as:

$$\hat{\tau} = \underbrace{\frac{\sum_i Y_i^{obs} W_i}{\sum_i W_i}}_{p} - \underbrace{\frac{\sum_i Y_i^{obs}(1 - W_i)}{\sum_i (1 - W_i)}}_{q}, \tag{6.3}$$

which represents the difference in the sample average outcome between the treated and untreated observations. For the splitting criterion, the gain difference after splitting is defined as:

$$D_{gain} = D_{after_{split}}(P^T, P^C) - D_{before_{split}}(P^T, P^C) \tag{6.4}$$

Where $D$ is the difference and $P^T$ and $P^C$ is the probability distribution of the outcome variable in the treatment and control groups (Rzepakowski & Jaroszewicz, 2012). The uplift trees were split using the Chi function, rooted in a statistical test that determines significant associations between two categorical variables. Within uplift modelling, this function aids in prioritising splits that highlight a significant relationship between the treatment and the outcome. The divergence in this method is represented by $X^2$:

$$X^2(P : Q) = \sum_{k=left,right} \frac{(p_k - q_l)^2}{q_k} \tag{6.5}$$

where $p$ indicates the sample mean in the treatment group, $q$ is the sample mean in the control group, and $k$ denotes the leaf in which $p$ and $q$ are calculated.

## 6.8. Additional Data and Analyses

This section presents additional data analysis elements not included in the main manuscript.

### 6.8.1. Correlations

Delegation behaviour exhibits weak positive correlations with STEM degrees, extraversion, agreeableness, conscientiousness, internal locus of control, and external locus of control. Conversely, it has weak negative correlations with gender (female), business and economics degrees, and neuroticism. Age and openness display almost no correlation with delegation behaviour (figure 6.8).

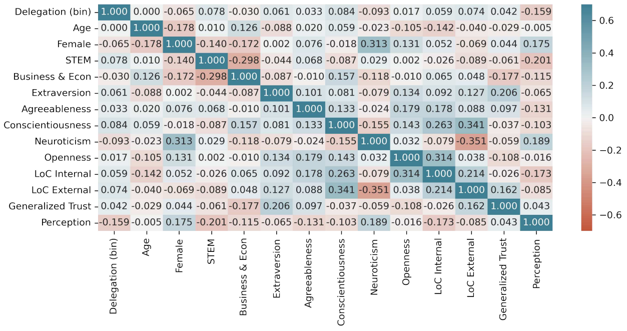


Figure 6.8.: Spearman correlation coefficients

Table 6.9 displays the results of point-biserial correlation coefficients between the personality traits and delegation behaviour (binary).

### 6.8.2. Regressions

This regression model includes interaction terms to account for the correlation between independent variables (table 6.10), providing a more nuanced analysis of the relationships between variables and delegation behaviour. In this model, the main effects of some variables change, and the added interaction terms help us better understand how the relationships between variables affect the outcome.

The internal locus of control variable becomes significant ($p = 0.041$) in the model with interaction terms, while it was not significant in the model without interactions. This change suggests that the relationship between internal locus of control and delegation behaviour

| Variable | Correlation Coefficient | p-value |
|---|---|---|
| Age | $-0.019^{*}$ | 0.026 |
| Female | $-0.065^{***}$ | $< 0.001$ |
| STEM | $0.078^{***}$ | $< 0.001$ |
| Business & Economics | $-0.03^{***}$ | $< 0.001$ |
| Extraversion | $0.06^{***}$ | $< 0.001$ |
| Agreeableness | $0.039^{***}$ | $< 0.001$ |
| Conscientiousness | $0.089^{***}$ | $< 0.001$ |
| Neuroticism | $-0.087^{***}$ | $< 0.001$ |
| Openness | $0.024^{**}$ | 0.005 |
| Internal LoC | $0.06^{***}$ | $< 0.001$ |
| External LoC | $0.068^{***}$ | $< 0.001$ |
| Generalised Trust | $0.044^{***}$ | $< 0.001$ |
| Perception | $-0.147^{***}$ | $< 0.001$ |

Table 6.9.: Point-biserial correlation coefficients to binary action of delegation

Significance levels: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

| Variable | Coefficient | Standard Error | p-value |
|---|---|---|---|
| Constant | 5.195 | 4.046 | 0.199 |
| Explanation | 0.195 | 0.21 | 0.354 |
| Payment | $-1.053^{***}$ | 0.24 | $< 0.001$ |
| Automation | 0.453 | 0.235 | 0.054 |
| Female | 0.821 | 0.567 | 0.147 |
| Age | $-0.012$ | 0.018 | 0.501 |
| STEM | 0.269 | 0.232 | 0.246 |
| Business & Economics | $-0.193$ | 0.201 | 0.337 |
| Extraversion | 0.018 | 0.059 | 0.755 |
| Agreeableness | 0.044 | 0.073 | 0.552 |
| Conscientiousness | $-0.411$ | 0.597 | 0.491 |
| Neuroticism | 0.111 | 0.361 | 0.759 |
| Openness | $-0.71$ | 0.416 | 0.088 |
| Internal LoC | $-1.266^{*}$ | 0.618 | 0.041 |
| External LoC | 0.251 | 0.604 | 0.678 |
| Generalised Trust | 0.07 | 0.065 | 0.284 |
| Perception | $-0.361^{*}$ | 0.14 | 0.01 |
| Female x Neuroticism | $-0.234$ | 0.136 | 0.084 |
| Internal LoC x Conscientiousness | 0.136 | 0.089 | 0.129 |
| External LoC x Conscientiousness | $-0.043$ | 0.094 | 0.65 |
| External LoC x Neuroticism | $-0.011$ | 0.079 | 0.894 |
| Internal LoC x Openness | 0.139 | 0.079 | 0.08 |

Table 6.10.: Logistic Regression results - delegation, with interaction Terms

Pseudo-$R^2 = 0.09$. Significance Levels: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

might be more complex than initially estimated by the first model. Including interaction terms allow us to capture the combined effects of internal locus of control with other variables, such as openness, which might help explain this shift in statistical significance.

The interaction between female gender and neuroticism is significant at the 10% level ($p = 0.084$). For instance, women generally report higher neuroticism scores than men (Costa Jr, Terracciano, & McCrae, 2001; Schmitt, Realo, Voracek, & Allik, 2008; Weisberg, DeYoung, & Hirsh, 2011), which is also true for our sample. Given that women generally report higher neuroticism scores than men, this term indicates that the relationship between neuroticism and delegation behaviour differs for males and females. Specifically, the effect of neuroticism on delegation behaviour may be more substantial for one gender than the other. As a result, the positive coefficient for the female gender in the second model suggests that the likelihood of delegation among females might depend more on their neuroticism level than males.

Another noteworthy interaction term is the one between internal locus of control and openness, which is significant at the 10% level ($p = 0.080$). This interaction suggests that the effect of internal locus of control on delegation behaviour is more pronounced for individuals with specific levels of openness. For example, participants with a high internal locus of control and high openness might be more likely to delegate tasks than those with a high internal locus of control and low openness. This finding further emphasises the importance of considering the interaction effects when examining the relationships between variables and delegation behaviour.

We also have fit quantile regression models (Koenker & Bassett Jr, 1978) using the cumulative frequency of delegation for each participant across all periods, removing the time dimension. We employed this method due to the varying relationships between the variables across different parts of the outcome distribution and the lack of normality. The results are summarised in table 6.11.

This model explains approximately 18.95% of the sample variance. Similarly to the logistic regression results, these findings show that the condition involving payment significantly reduces the frequency of delegation ($p < 0.001$), while full automation significantly increases it ($p = 0.002$). Among personal characteristics, only External Locus of Control significantly contributes to delegation, indicating that participants who believe outcomes are beyond their control are more likely to delegate decisions ($p = 0.04$). Moreover, a negative perception of algorithms significantly corresponds to a less frequent delegation of decisions ($p < 0.001$). Other actors such as explanation condition, demographics, Big Five personality traits, Internal Locus of Control, and Trust do not significantly affect the delegation frequency. We have also controlled for correlated variables in this model by adding interaction terms; the results are summarised in table 6.12.

Upon adding interaction terms, the pseudo-R-squared value rose to 21.01%, showing a marginally improved model fit. Payment ($p < 0.001$) and automation ($p = 0.01$) still significantly influence delegation. Notably, individuals with a STEM background ($p = 0.017$) show a significant positive association with delegation. Openness to experience negatively correlates with delegation ($p = 0.034$). A significant interaction emerges between internal locus of control and Openness ($p = 0.04$): those high in internal locus of control and openness tend to delegate more. A negative view of algorithms remains a strong deterrent to delegation ($p = 0.002$).

### 6.8.3. Machine Learning

In the main text, we illustrate the Machine Learning models' feature impacts on the predictions with a combined plot. Each model's separate feature importance values are displayed in figure 6.9.

| Variable | Coefficient | Standard Error | p-value |
|---|---|---|---|
| Intercept | 0.386 | 0.28 | 0.169 |
| Explanation | 0.088 | 0.059 | 0.138 |
| Payment | $-0.367^{***}$ | 0.06 | $< 0.001$ |
| Automation | $0.192^{**}$ | 0.06 | 0.002 |
| Female | $-0.027$ | 0.046 | 0.558 |
| Age | $-0.003$ | 0.005 | 0.526 |
| STEM | 0.078 | 0.057 | 0.173 |
| Business & Economics | $-0.051$ | 0.055 | 0.348 |
| Extraversion | 0.023 | 0.017 | 0.179 |
| Agreeableness | 0.011 | 0.021 | 0.598 |
| Conscientiousness | 0.025 | 0.024 | 0.285 |
| Neuroticism | 0.002 | 0.018 | 0.927 |
| Openness | 0.005 | 0.023 | 0.841 |
| Internal LoC | $-0.02$ | 0.027 | 0.467 |
| External LoC | $0.058^{*}$ | 0.028 | 0.04 |
| Generalised Trust | 0.013 | 0.019 | 0.49 |
| Perception | $-0.143^{***}$ | 0.034 | $< 0.001$ |

Table 6.11.: Quantile Regression results - cumulative delegation frequencies

Pseudo-Pseudo-$R^2 = 0.19$. Significance Levels: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

| Variable | Coefficient | Standard Error | p-value |
|---|---|---|---|
| Intercept | 1.206 | 1.043 | 0.248 |
| Explanation | 0.039 | 0.057 | 0.497 |
| Payment | $-0.397^{***}$ | 0.058 | $< 0.001$ |
| Automation | $0.151^{*}$ | 0.058 | 0.01 |
| Female | 0.105 | 0.139 | 0.45 |
| Age | $-0.007$ | 0.005 | 0.14 |
| STEM | $0.131^{*}$ | 0.055 | 0.017 |
| Business & Economics | $-0.034$ | 0.052 | 0.512 |
| Extraversion | 0.011 | 0.017 | 0.509 |
| Agreeableness | 0.015 | 0.02 | 0.451 |
| Conscientiousness | 0.058 | 0.166 | 0.727 |
| Neuroticism | 0.02 | 0.086 | 0.814 |
| Openness | $-0.22^{*}$ | 0.103 | 0.034 |
| Internal LoC | $-0.258$ | 0.144 | 0.074 |
| External LoC | 0.18 | 0.157 | 0.253 |
| Generalised Trust | 0.008 | 0.018 | 0.672 |
| Perception | $-0.102^{**}$ | 0.033 | 0.002 |
| Female x Neuroticism | $-0.032$ | 0.033 | 0.331 |
| Internal LoC x Conscientiousness | 0.016 | 0.026 | 0.548 |
| External LoC x Conscientiousness | $-0.025$ | 0.025 | 0.307 |
| External LoC x Neuroticism | $-0.003$ | 0.018 | 0.879 |
| Internal LoC x Openness | $0.041^{*}$ | 0.02 | 0.04 |

Table 6.12.: Quantile Regression Results - cumulative delegation frequencies, with interaction terms

Pseudo-$R^2 = 0.21$. Significance Levels: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

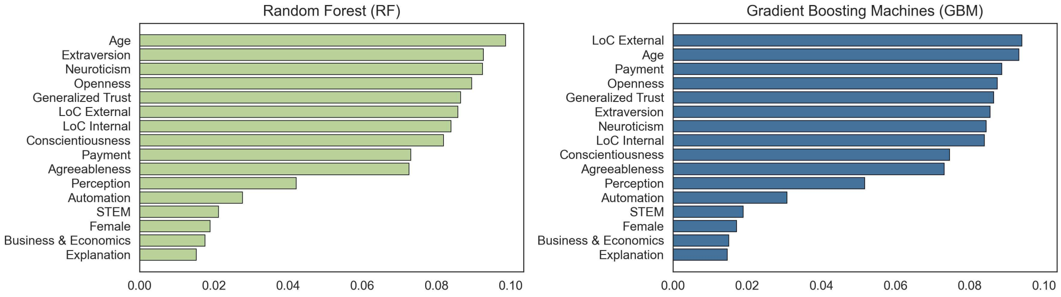


Figure 6.9.: Feature Importances for RF and GBM

Overall, both models show similar feature importance quantifications. The main differences lie in the relative importance of LoC External, Payment, and certain personality traits like Extraversion and Neuroticism. GBM emphasises external locus of control and payment more, while RF highlights age and personality traits.

## 6.9. Technical Remarks

The documented experiment was executed online, programmed with the oTree open-source platform (D. L. Chen et al., 2016). The data work was performed using Python language. The statistical tests were done using statsmodels (Seabold & Perktold, 2010). The machine learning models were deployed, tuned, and cross-validated using Scikit-Learn (Pedregosa et al., 2011). Both models were tuned using a grid search algorithm with the target to maximise the AUC-ROC. It is important to outline that this is a computationally expensive procedure. The parameter set for the Random Forest model is in table 6.13.

Similarly, the grid search-generated parameters for the GBM model are described in table 6.14

| Parameter | Value | Definition |
|---|---|---|
| bootstrap | True | Determines whether or not to use bootstrap samples when building trees |
| class_weight | balanced_subsample | Adjusts the weights of the classes. balanced_subsample means it computes weights based on the bootstrap sample for every tree |
| criterion | entropy | Defines the function to measure the quality of a split. entropy is for information gain |
| max_depth | 15 | Specifies the maximum depth of the tree |
| max_features | sqrr | The number of features to consider when looking for the best split. sqrt means the square root of the total number of features |
| min_samples_leaf | 1 | The minimum number of samples required to be at a leaf node |
| min_samples_split | min_samples_split | The minimum number of samples required to split an internal node |
| n_estimators | 100 | The number of trees in the forest |

Table 6.13.: Random Forest Classifier parameters

| Parameter | Value | Definition |
|---|---|---|
| learning_rate | 0.05 | Determines the impact of each tree on the final outcome |
| max_depth | 10 | Specifies the maximum depth of the tree |
| max_features | sqrt | The number of features to consider when looking for the best split. sqrt means the square root of the total number of features |
| min_samples_leaf | 1 | The minimum number of samples required to be at a leaf node |
| min_samples_split | 15 | The minimum number of samples required to split an internal node |
| n_estimators | 100 | The number of boosting stages to perform. Each stage adds a new tree into the ensemble |
| subsample | 0.7 | The fraction of samples to be used for fitting the individual base learners |

Table 6.14.: Gradient Boosting Machine Classifier parameters

The cross-validation technique used in both models was the GroupKFold algorithm, which aggregated samples for the same participant. This procedure was performed in both the parameter search and model training steps, using five validation folds. In this process, we split the training data into a number of subsets or "folds." We train the model on the remaining data for each fold and test it on this fold. This process is repeated for each fold, allowing us to assess the model's performance based on its ability to predict new data (Berrar, 2019; Kohavi et al., 1995). With an equivalent objective as clustering the regression errors on a participant level (section 6.5.3), we aggregated the participant observations here using the GroupKFold variant, which ensures instances from the same participant either in the training set or the test set. This approach safeguards against data leakage and maintains a realistic estimate of the model's performance, especially when observations within the same group (in this case, participant) are correlated.

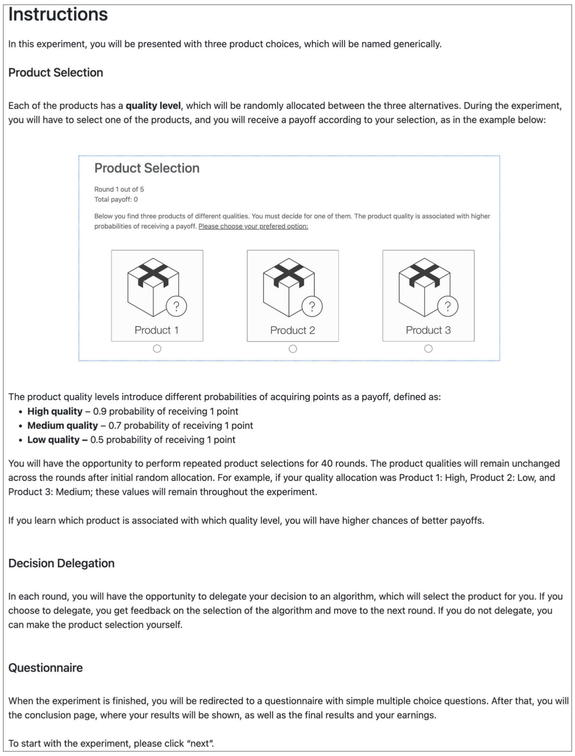
The uplift random forest classifier was implemented using the causalml library (H. Chen et al., 2020). Since this method, in conjunction with the group cross-validation using synthetic control groups, was performance costly, we implemented a less-exhaustive approach for the parameter-fitting method, using the Optuna library (Akiba, Sano, Yanase, Ohta, & Koyama, 2019). It employs efficient search algorithms, such as Tree-structured Parzen Estimator (TPE). We ran an optimisation study for 150 trials and selected the parameter set that yielded satisfactory AUUC scores. One important remark here is that calculating the AUUC in this way might produce abnormally high results due to the stochastics in place, so practitioners might have to supervise the optimisation process. Table 6.15 describes the parameter values.

| Parameter | Value | Definition |
|---|---|---|
| n_estimators | 850 | The number of trees in the forest |
| max_depth | 8 | The maximum depth of each decision tree |
| max_features | 9 | The number of features to consider when looking for the best split |
| min_samples_leaf | 45 | The minimum number of samples required to be at a leaf node |
| min_samples_treatment | 15 | The minimum number of samples in a leaf node that come from the treatment group |
| n_reg | 14 | The regularization parameter used in the causal tree procedure |
| evaluationFunction | Chi | The evaluation function used to evaluate splits |

Table 6.15.: Uplift Random Forest Classifier parameters

## 6.10. Experiment Design Screens

In this appendix session, we added the most important screens for the experiment. Figure 6.10 displays the instructions screen. Figure 6.11 contains the main task screens for each treatment. Figure 6.12 shows the attention questions.



Instructions

In this experiment, you will be presented with three product choices, which will be named generically.

Product Selection

Each of the products has a **quality level**, which will be randomly allocated between the three alternatives. During the experiment, you will have to select one of the products, and you will receive a payoff according to your selection, as in the example below:

Product Selection

Round 1 out of 5
Total payoff: 0

Below you find three products of different qualities. You must decide for one of them. The product quality is associated with higher probabilities of receiving a payoff. Please choose your prefered option:

The product quality levels introduce different probabilities of acquiring points as a payoff, defined as:

- **High quality** – 0.9 probability of receiving 1 point
- **Medium quality** – 0.7 probability of receiving 1 point
- **Low quality –** 0.5 probability of receiving 1 point

You will have the opportunity to perform repeated product selections for 40 rounds. The product qualities will remain unchanged across the rounds after initial random allocation. For example, if your quality allocation was Product 1: High, Product 2: Low, and Product 3: Medium; these values will remain throughout the experiment.

If you learn which product is associated with which quality level, you will have higher chances of better payoffs.

Decision Delegation

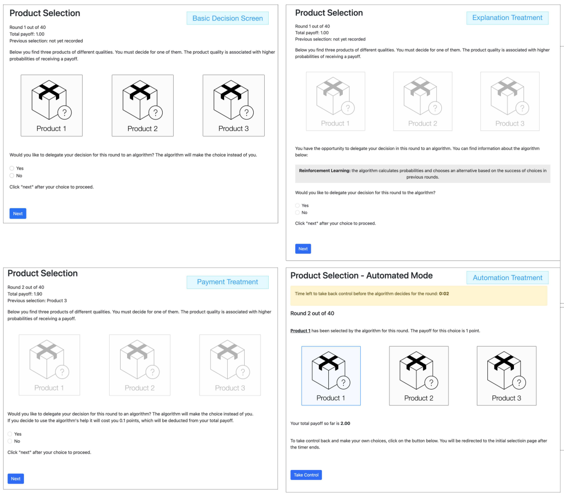
In each round, you will have the opportunity to delegate your decision to an algorithm, which will select the product for you. If you choose to delegate, you get feedback on the selection of the algorithm and move to the next round. If you do not delegate, you can make the product selection yourself.

Questionnaire

When the experiment is finished, you will be redirected to a questionnaire with simple multiple choice questions. After that, you will the conclusion page, where your results will be shown, as well as the final results and your earnings.

To start with the experiment, please click "next".

Figure 6.10.: General Instructions Screen



Product Selection

Round 1 out of 40
Total payoff: 1.00
Previous selection: not yet recorded

Below you find three products of different qualities. You must decide for one of them. The product quality is associated with higher probabilities of receiving a payoff.

Would you like to delegate your decision for this round to an algorithm? The algorithm will make the choice instead of you.

○ Yes
○ No

Click "next" after your choice to proceed.

Product Selection

Round 1 out of 40
Total payoff: 1.00
Previous selection: not yet recorded

Below you find three products of different qualities. You must decide for one of them. The product quality is associated with higher probabilities of receiving a payoff.

You have the opportunity to delegate your decision in this round to an algorithm. You can find information about the algorithm below:

**Reinforcement Learning:** the algorithm calculates probabilities and chooses an alternative based on the success of choices in previous rounds.

Would you like to delegate your decision for this round to the algorithm?

○ Yes
○ No

Click "next" after your choice to proceed.

Product Selection

Round 2 out of 40
Total payoff: 1.90
Previous selection: Product 3

Below you find three products of different qualities. You must decide for one of them. The product quality is associated with higher probabilities of receiving a payoff.

Would you like to delegate your decision for this round to an algorithm? The algorithm will make the choice instead of you.
If you decide to use the algorithm's help it will cost you 0.1 points, which will be deducted from your total payoff.

○ Yes
○ No

Click "next" after your choice to proceed.

Product Selection - Automated Mode

Time left to take back control before the algorithm decides for the round: **0:02**

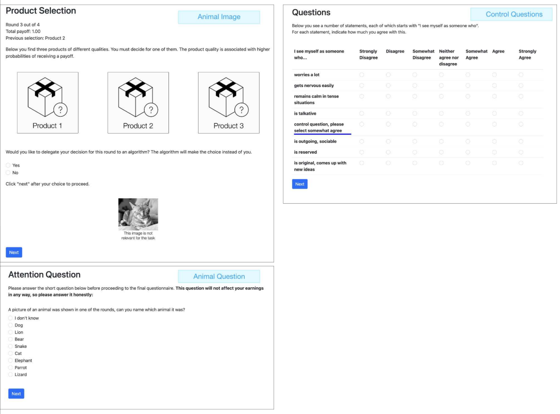
Round 2 out of 40

**Product 1** has been selected by the algorithm for this round. The payoff for this choice is 1 point.

Your total payoff so far is **2.00**

To take control back and make your own choices, click on the button below. You will be redirected to the initial selectioin page after the timer ends.

Figure 6.11.: Main Task Experiment Screens



Product Selection

Round 3 out of 4
Total payoff: 1.00
Previous selection: Product 2

Below you find three products of different qualities. You must decide for one of them. The product quality is associated with higher probabilities of receiving a payoff.

Would you like to delegate your decision for this round to an algorithm? The algorithm will make the choice instead of you.

○ Yes
○ No

Click "next" after your choice to proceed.

Questions

Below you see a number of statements, each of which starts with "I see myself as someone who".
For each statement, indicate how much you agree with this.

| I see myself as someone who... | Strongly Disagree | Disagree | Somewhat Disagree | Neither agree nor disagree | Somewhat Agree | Agree | Strongly Agree |
|---|---|---|---|---|---|---|---|
| worries a lot | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| gets nervous easily | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| remains calm in tense situations | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| is talkative | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| control question, please select somewhat agree | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| is outgoing, sociable | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| is reserved | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| is original, comes up with new ideas | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

Attention Question

Please answer the short question below before proceeding to the final questionnaire. **This question will not affect your earnings in any way, so please answer it honestly:**

A picture of an animal was shown in one of the rounds, can you name which animal it was?

○ I don't know
○ Dog
○ Lion
○ Bear
○ Snake
○ Cat
○ Elephant
○ Parrot
○ Lizard

Figure 6.12.: Attention Measures

# 7. When To Get That Extra Paycheck? A Behavioural Evaluation Of A Default Change

## Authors

Leon Houf, Benedikt Vogt & Rik Dillingh [1]

## Abstract

Choice architecture has been a prominent theme in recent literature where special interest has been paid to nudging. We study a situation where an existing nudge is partially revoked and structured as a more open choice architecture. For this, we exploit a natural experiment in the Netherlands where civil servants, through a new policy, can decide when and how to get their $13^{th}$ and $14^{th}$ salary payments within a year. Each level of government implemented the policy at different points in time with different defaults on the same initial choice set. We use administrative panel data on more than 21 million monthly paychecks. We investigate how a default with an immediate payout and a default with a delayed payout influence choices, respectively. First, we find that in the first year of treatment, there is less than 2% difference in deviation between the two defaults. Second, the longer the defaults are in place, stickiness to the default with an immediate payout is about 15% higher. Third, we find that for civil servants with no experience with the old system, the default stickiness in the immediate payout default is an additional 13% higher.

## Keywords

Choice Architecture, Defaults, Natural Experiment, Policy Evaluation, Administrative Data

## 7.1. Introduction

Choice architecture and defaults influence individual choices in various settings. The recent literature shows that defaults determine important decisions in life, such as retirement savings (Madrian & Shea, 2001), charity donations (Altmann, Falk, Heidhues, Jayaraman, & Teirlinck, 2019) or organ donations (Johnson & Goldstein, 2003).

The current literature remains silent on the way defaults affect how individuals receive their salary. However, this is important because both theoretical (Parsons & Van Wesep, 2013) and recent empirical evidence (Brune, Chyn, & Kerwin, 2021) suggest that individuals exhibit preferences for deferred payment of their salaries. For individuals who

[1] **Status:** This paper was presented at an CPB Expert Session in November 2021, at Heidelberg University in January 2022, at ASFEE in Lyon in May 2022 and at M-BEPS in Maastricht in June 2024. Currently, we prepare a submission to a journal such as *Behavioural Public Policy.*

do not have self-control problems, such paternalistic payout schemes are not necessary to provide them with sufficient liquidity and funds for their consumption and savings decisions. However, individuals with self-control problems can have a preference for delayed payouts (Laibson, 1997; Parsons & Van Wesep, 2013). One of the main mechanisms for such deferred payments are commitment devices in the form of illiquid (mental) accounts to help individuals with self-control problems build up buffers for greater expenditures.

In this paper, we study the effect of a change in choice architecture on the payout patterns of individual salaries. First, we investigate how many individuals adhere to a new payout default and how persistent this behaviour is over time. Second, we investigate if the type of default matters for adherence. Third, we investigate if the experience with an earlier default influences adherence to the new default.

We obtain four main findings. First, the majority of individuals deviate from a newly set default on the payout pattern of their salary. Second, default adherence depends on the type of default. The greater the difference from the default after the policy change compared to the initial situation, the higher the probability to deviate. In the new situation, when a big deferred payment is a new default, 60% of the population deviate from the default, whereas in the situation which has no deferred payment, 53% deviate. Third, deviation from the default decreases over time. Fourth, deviation from the default depends on the experience with the old system. Individuals who had little experience with the old system are 14% (8.6%-points) more likely to deviate from the default in the system with big deferred payments.

In many decision-making settings, some default choice has to be made by an institution, government or firm. We show that defaults are among the most significant determinants, also in important cases like the payout patterns of monthly salaries. Even if the deliberate goal of a policy was to give more freedom of choice in the timing of salary payments, the majority of individuals do not make an active decision and simply stick to the default presented. This illustrates the importance of a careful default choice of the responsible institution at stake because payout patterns can have implications for individual consumption and savings.

We make use of a natural experiment to identify the causal effects of introducing two alternative types of defaults on the payout patterns of individual salaries. In our setting a well-established system as in many countries (Parsons & Van Wesep, 2013) of receiving an extra salary before summer and another extra salary before the Christmas holidays was abolished. These ‘double pay-check’ days had a paternalistic reasoning in helping workers save for these events.

We make use of full population administrative data based on all payrolls of the public sector in the Netherlands in the 2012 to 2021 period. We first identify revealed default behaviour in the payroll patterns and compare then if the type of default and the period of the default influence how individuals adapt their salary payout preferences.

We make the following contributions to the current literature on defaults in economics and psychology. We contribute to the literature on the effects of defaults, their effects on choices and persistence. Our first contribution lies in new evidence on the role of defaults regarding salary payout patterns. Whereas many studies have shown the importance of defaults in other settings (Johnson & Goldstein, 2003; Madrian & Shea, 2001), our evidence is new. Our second contribution lies in showing the development of default adherence over time and the heterogeneity between different types of defaults. Recent papers (Altmann et al., 2019; Heidhues & Strack, 2019) show that the default type matters for how much money people donate. We show that the type of default also matters in other settings. Third, we also show that experience with a previous default affects the extent to which people deviate from newly set defaults.

The remainder of the paper is structured as follows. In section 7.2 we explain the institutional setting and the policy changes we investigate. In section 7.3 we briefly describe the data and then present our results in section 7.4. We conclude in section 7.5.

## 7.2. The Policy Change

In this section, we describe the institutional setting. We first describe the setting before the policy change was introduced and then introduce the two main changes in defaults and the differences in timing between governmental layers.

### 7.2.1. Before Policy

For a long time, civil servants (and many workers in the private sector) in the Netherlands received their total yearly salary in 14 almost equal parts (before tax), of which twelve are paid as monthly payments. The remaining two parts were paid out as an unconditional, additional and fixed payment. The $13^{th}$ part was labelled “vacation allowance” and paid out with the monthly salary in May as a double paycheck day. The $14^{th}$ part was labelled “end of year payment” and was paid towards the end of the year; in most governmental layers in November, for some in December, together with the monthly salary.

The accumulation cycle for these extra payments works such that a worker earns $2 \times \frac{1}{12}$ of their basic monthly salary extra in every month until the accumulated sum is paid out in May or the end of the year, respectively. Then, the accumulation cycle starts again. It is important to note that the employer automatically defers these payments in this system but does not withhold them for an earlier payout, because in principle, within such a cycle, it was possible for workers to have a payout of the already accumulated salary in this cycle. Therefore, the choice set of the workers always includes the options for earlier payment or sticking to the default of receiving the payment at the end of the payment cycle labelled as vacation allowance or end-of-year payment.

This regulated, special timing of pay can be understood through the model of Parsons and Van Wesep (2013) as an attempt to help the workers mitigate their self-control problems of spending money too early. This is based on the assumption that under perfect self-control, the workers would prefer to spend extra in summer and at the end of the year, for example, on vacation. Therefore, the employer withholds a part of the total annual salary (in our case $\frac{2}{14} \approx 14.3\%$ of the total yearly salary) and pays it out at the time the workers supposedly benefit from it the most.

The paternalistic nature of this payment scheme is further shown through labelling the additional May payment "vacation allowance". This can be understood as an explicit spending recommendation by the employer, which for our sample of civil servants can be understood as a government recommendation. Abeler and Marklein (2017); Kooreman (2000); Thaler (1990) show that labelling of (mental) accounts indeed does change the propensity to spend and consume a certain product or category of goods. It is important to note that the timing schedule we observe has been stable for a long time. It is not actually clear whether workers benefit from the additional payments in May and the end of the year or whether a different payment schedule would be preferable for some workers. Nibud (2019) surveyed civil servants about the use of their "vacation allowance" and found that 43% of respondents use it for vacation, 38% for savings and only 13% for household expenses. This suggests that for many civil servants, the nudge worked as intended.

### 7.2.2. The Policy Change

The presented policy of delayed payments can maximise the utility of employees under the assumption that civil servants have a higher utility when receiving $\frac{1}{14}$ of their total yearly salary both towards the middle and the end of the year, as described in the model of Parsons and Van Wesep (2013). While many people in the survey by Nibud (2019) indicate that they prefer this situation to a flat monthly payment, it is not clear that this nudge is optimal as a default for everyone.

This assumption is withdrawn in the policy change we investigate. The employer introduced this policy to 'enhance the control of the own career and make choices for work-life balance'.[2] The new policy transforms the $13^{th}$ and $14^{th}$ payment part into an "Individual Choice Budget" (ICB) while the basic monthly payments remain unchanged.

Therefore the goal of the policy change is to explicitly not presume what the ideal timing of pay of the individual choice budget is for a given worker. The new policy's design revokes the previous policy's strong nudge on the timing of pay and, in a sense, "loosens the ties" that strap Odysseus to the mast. In the following section we will show how the choice set of the new policy is largely equivalent to the old policy and how its choice architecture leads to a 'de-nudging', i.e. not having a targeted decision outcome in the pay timing.

Equivalent to the previous policy, under the new policy, a worker earns $\approx 2 \times \frac{1}{12}$ of their basic monthly salary in every month as this budget. This is the same sum as before, so the total yearly salary stays constant. The budget sums up to $12 * \frac{2}{12} = 2$ salary parts, mirroring the $13^{th}$ and $14^{th}$ payment from before. [3] One crucial difference of the new policy is the cycle of accumulating this budget. In the previous system, there were, in

[2] See https://www.caorijk.nl/cao-rijk/hoofdstuk-9/individueel-keuzebudget-ikb for more information.

[3] In the first year $t$ of the introduction of the policy, $\frac{7}{12}$ of the cycle of the vacation allowance was still paid out in May because this was already built up from June to December the previous year $t-1$. This one-off payment at the change of the system increases the total payment in the year of introduction $t$ by $\frac{7/12}{14parts} \approx 4.2\%$.

fact, two cycles. One ends with the payment of the vacation allowance in the middle of the year and then starts again for 12 months, and the other one ends with the payment at the end of the year and then starts again. The ICB accumulation cycle starts in January and ends in December. At any point within this cycle, workers can use the already accumulated budget and indicate their choices through an online HR platform.

Employees can use their budget in several ways. First, the online platform allows to indicate a monthly payment of the ICB as soon as it is earned. Next to this, workers can decide at any point to have a payment up to the already earned amount. This means that it is also possible to mirror the payment pattern of the previous policy with a payout in the middle of the year and towards the end of the year. [4] At the end of a cycle, i.e. at the end of a calendar year, the unused budget is paid out automatically together with the December salary.

### 7.2.3. Setting of the Natural Experiment

In line with the policy's goal, the new system does not require workers to make a choice. This system does not enforce a choice, but since there is a choice set, a default is necessary to put in place. This section first introduces the two different types of defaults that have been introduced at different government layers, and how we can use this natural experiment.

#### 7.2.3.1. The Monthly Default Treatment

In the Monthly Default Treatment, the ICB is paid out as soon as it is earned together with the basic monthly salary. For a worker who does not make a change, this results into a flat payment schedule within the year and a $\frac{2}{12} \approx 16.7\%$ increased monthly payout compared to the basic monthly salary before the policy change, while the total yearly salary stays the same. If a worker decides to actively change the payment structure to receive the ICB later, this can be interpreted as using the employer as a savings account.[5] At the latest, the ICB is paid out in December. In the framework of Laibson (1997), this can be explained as rational behaviour for agents who want to spend the money later within the year and know that they have self-control problems if they receive the money earlier than their intended consumption day. To commit themselves to the later consumption time, they delay their own payment. The monthly default is used by the provinces for all their workers. [6]

#### 7.2.3.2. The December Default Treatment

The December Default Treatment pays the entire ICB to the worker in December. If a worker has not used anything of the ICB during the year, the final paycheck in December

[4] Furthermore, and as an extension of the previous policy, the ICB can also be used for non-monetary payments. Workers can buy "ICB-hours" for the price of the actual hourly wage through this budget until a limit. These hours can be used to reduce the working hours within a year. Still, they can also be saved up for a sabbatical or early retirement.

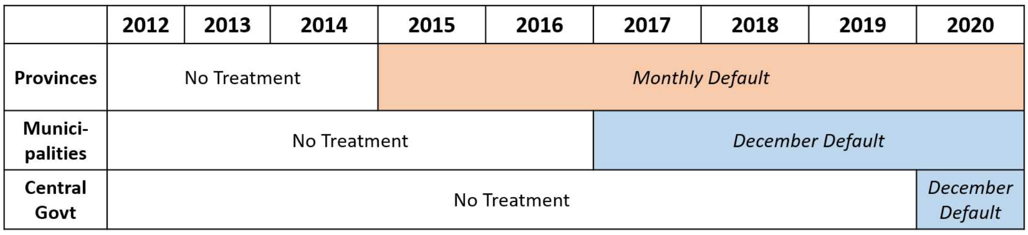
[5] This is an equivalent to a savings account with zero interest.

[6] Not all provinces introduced the default in 2015. Of the overall twelve provinces, we exclude the four provinces which introduced the ICB in 2016 because of their small sample size in the further analysis.

would then include the usual monthly salary ($\frac{1}{14}^{th}$ of the total yearly salary) and the ICB ($\frac{2}{14}^{th}$ of the total yearly salary), so in total $\frac{3}{14}^{th}$ of the gross total yearly salary. If a worker decides to use parts of the ICB earlier than December, only the remaining part is paid out in December.

#### 7.2.3.3. Natural Experiment

Figure 7.1 shows the differences in timing and defaults between the governmental layers. Here we observe the natural experiment that governmental layers differ in the option of the default choice set. There is no specific (endogenous) reason why a given governmental layer chooses the respective default. Therefore we can analyse this layered introduction as a natural experiment. We observe three governmental layers introducing the ICB with either the monthly or December default. The ICB with monthly default was introduced by most provinces for their workers in 2015. The workers of the municipalities were introduced to the ICB with the December default in 2017, the central government workers in 2020.



| | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 |
|---|---|---|---|---|---|---|---|---|---|
| **Provinces** | No Treatment | | | *Monthly Default* | | | | | |
| **Munici-palities** | No Treatment | | | | | *December Default* | | | |
| **Central Govt** | No Treatment | | | | | | | | *December Default* |

Figure 7.1.: Timing of Default Changes

*All governmental layers started with the similar previous system. The provinces in our sample introduced the Monthly Default Treatment in 2015. The municipalities introduced the December Default Treatment in 2017 and the central government in 2020.*

## 7.3. Data

We use administrative panel data from Statistics Netherlands (Centraal Bureau voor Statistiek, CBS). These data allow us to observe monthly payment data from all civil servants by their employers. We combine these observations to data on household composition and household wealth.

### 7.3.1. Descriptive Statistics Full Sample

The individuals in the different default regimes are comparable regarding the most important socioeconomic characteristics. Table 7.1 shows descriptive statistics about the household balance sheet, yearly income and personal and household characteristics. The number of observations is substantially smaller for the monthly treatment than for the December treatment, because only workers in the Provinces were presented with the monthly default and the Provinces constitute the smallest of the governmental layers in the Netherlands. The net available household income, the age and the household size and composition are very similar between both treatment groups. Workers in the provinces tend to be more wealthy on average and the share of women is smaller, but the differences are limited and overall the groups are quite comparable.

| | Monthly Treatment | | December Treatment | |
|---|---|---|---|---|
| | Mean | Median | Mean | Median |
| Age | 49.0 | 51 | 48.2 | 50 |
| Share of Women | 44% | | 48% | |
| Number of Adults in Household | 1.8 | 2 | 1.8 | 2 |
| Number of Children in Household | 0.9 | 1 | 0.9 | 1 |
| Gross Indiv Year Income (T€) | 57.4 | 56 | 50.4 | 48 |
| Net available HH income (T€) | 63.6 | 62 | 63.2 | 61 |
| Bank Account Balance (T€) | 61.2 | 35 | 50.1 | 28 |
| Financial Assets (T€) | 75.5 | 40 | 58.4 | 31 |
| Total Wealth (T€) | 210.3 | 145 | 189.8 | 126 |
| N | 9,453 | | 308,845 | |

Table 7.1.: Descriptive Statistics of the Full Sample

*Mean and Median over 2012 to 2021 of full samples split by which default system the employer used.*

### 7.3.2. Comparison Experienced and New Civil Servants

Table 7.2 compares the sample split by their experience. Experienced civil servants are those who have experienced the previous payment system - with a double paycheck before the summer and at the end of year - for at least 5 years in their current job, before the introduction of the ICB. This is clearly visible in the descriptive statistics for both groups. New civil servants - who constitute about 8 percent of the sample - are on average substantially younger and also have a lower household income and wealth. Also, the share of women in more recent cohorts of the civil service has increased substantially.

| | Experienced Civil Servants | | New Civil Servants | |
|---|---|---|---|---|
| | Mean | Median | Mean | Median |
| Age | 52.4 | 54 | 26.4 | 27 |
| Share of Women | 46% | | 55% | |
| Number of Adults in Household | 1.8 | 2 | 1.8 | 2 |
| Number of Children in Household | 0.9 | 1 | 0.6 | 1 |
| Gross Indiv Year Income (T€) | 54.0 | 52 | 35.3 | 38 |
| Net available HH income (T€) | 64.1 | 62 | 58.1 | 55 |
| Bank Account Balance (T€) | 55.6 | 32 | 31.4 | 17 |
| Financial Assets (T€) | 65.2 | 35 | 34.2 | 18 |
| Total Wealth (T€) | 207.3 | 148 | 98.4 | 34 |
| N | 208,470 | | 17,331 | |

Table 7.2.: Descriptive Statistics of the Samples Split by Experience

### 7.3.3. Identification of Choice Behaviour

We do not observe the actual choices of the employees in the HR system, but we observe the revealed choices as monthly payroll patterns of every civil servant in a year. This allows us to infer the actions a civil servant took.

We categorise these actions as (i) *monthly*: having a monthly, flat payment pattern, (Default in the Monthly Default Treatment) (ii) *December*: receiving only the basic salary during the year and having the entire ICB paid out in December, (Default in the December

Default Treatment) (iii) *pre-system*: imitating the pre-system payment schedule with one payout towards the middle of the year and one towards the end of the year and (iv) *mixed*: as mixtures of these that do not clearly fall into any of these categories.

We categorise the observed salary payments using the following steps. In December, the remaining ICB budget is paid out regardless of the default. Therefore, the number of *Salary Parts* in December can be used to analyse the default sticking behaviour. Figure 7.2 shows how different values of *Salary Parts* paid in December correspond to the different outcomes depending on the default. The left side shows the number of *Salary Parts* the treatment groups received in December in the year before the ICB policy was introduced (between the 25th and 75th percentile). For the Central Government, that is close to one, given that they used to receive their double paycheck in November. For the other treatment groups, who used to receive their double paycheck in December, the number of *Salary Parts* they received in December in the year before the introduction of the ICB polity was between 1.6 and 2.0. The right side categorises the situation after the policy introduction.

The cutoffs for the categories are robust against noise. For untreated subjects in months other than May, November or December, the paid *Salary Parts* should be close to 1 and not be above 1.4. This holds indeed for 98.46% of these observations, and only 0.14% observations receive *Salary Parts* of more than 2.8 when it should be close to 1.

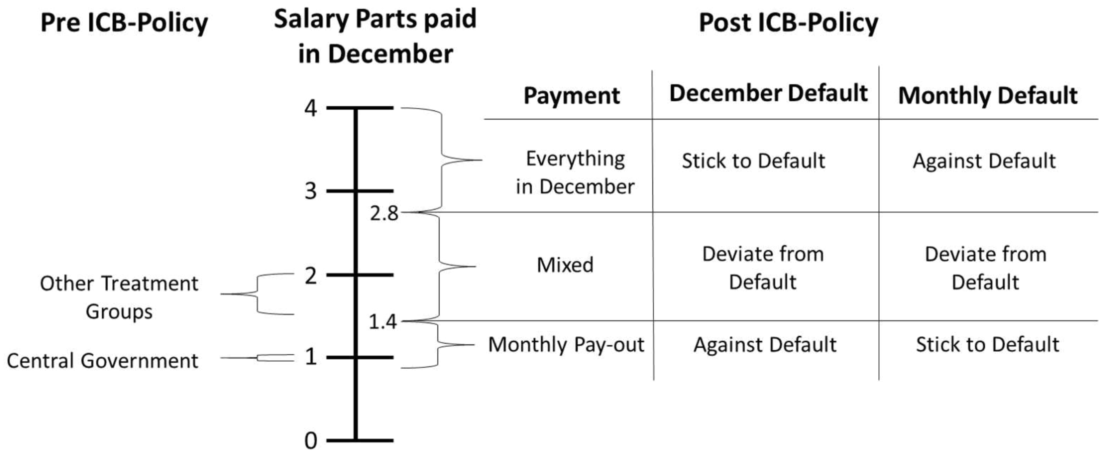


Figure 7.2.: Overview Default and Payment Patterns in December

- *Salary Parts* $\geq 2.8$ represents a situation where a subject receives the complete ICB budget in December. This outcome will be labelled *Outcome* =“December Payment“. This outcome is the default for subjects in the December default treatment who need not act to receive this outcome. Subjects in the monthly default treatment need to take action to receive it.
- $1.4 \leq$ *Salary Parts* $< 2.8$ shows a mixed payment and will therefore be labelled *Outcome* = “Mixed“. For both defaults, this needs an action by the subject. This

action does not systematically change the payment pattern indicated by the default, yet it still is a deviation.

- *Salary Parts* $< 1.4$ represents a situation where a subject receives the ICB budget in monthly payments. This outcome will be labelled *Outcome* =“Monthly Payment“. To have this payment pattern requires an action in the December default treatment and no action in the monthly default treatment.

While figure 7.2 presents the underlying concept of how to interpret the distribution of *Salary Parts*, figure 7.3 shows on the left side the distribution patterns in December in the year before and on the right side in the year with the policy introduction for Municipalities (December default) and Province ’15 (Monthly Default).

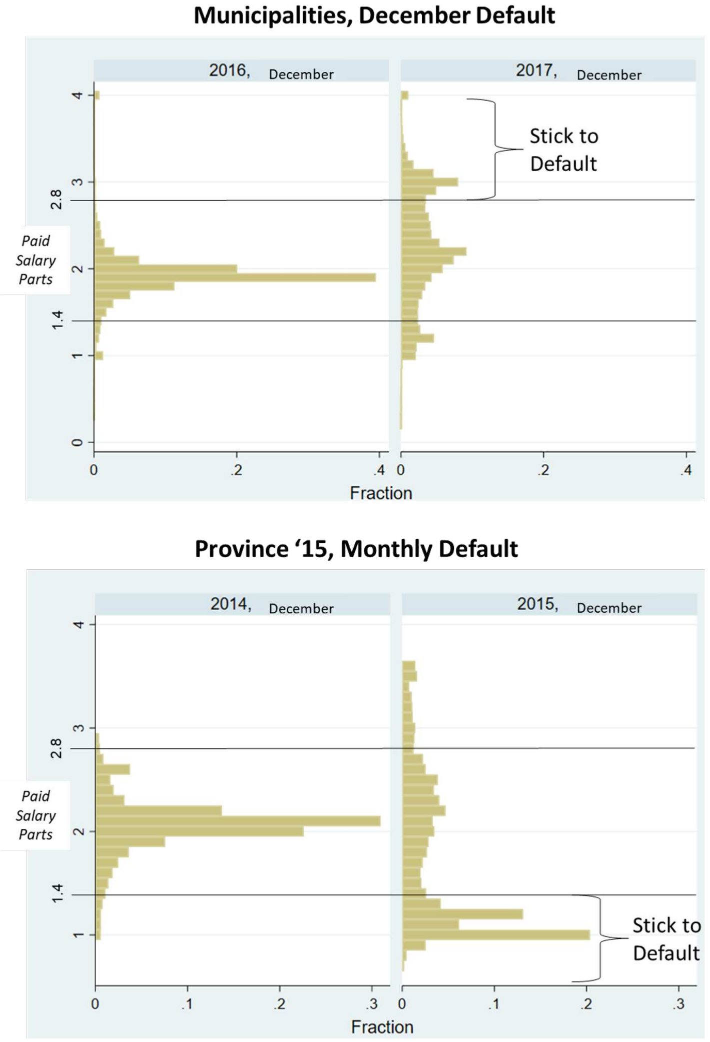


Figure 7.3.: Paid Salary Parts in December before and after Policy Introduction

The distribution before the policy is similar for both groups, with a normal distribution with a mean close to 2 and a relatively small standard deviation. After the policy, however, the patterns are clearly distinct. In both treatments, the received *Salary Parts* are scattered between 1 and 3.5. In both treatments, there is a small peak close to 2.2, which could represent subjects that try to mimic the pre-policy situation, which is a deviation from either of the new ICB defaults. In the monthly default treatment, there are huge peaks at 1 and 1.2, both representing situations where subjects stick to the default. In the December default treatment, there is a peak close to 3, which represents subjects that stick to the default.

## 7.4. Results

### 7.4.1. Deviating from default over time differs with experience of pre-system and type of default

We first analyse the deviation behaviour of civil servants who have experience with the previous system. This we compare to the civil servants that are new to the system.

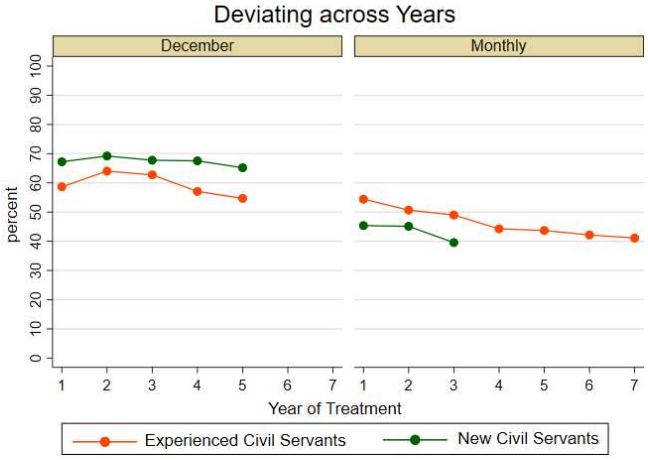


Figure 7.4.: Deviation across treatments

The orange line in figure 7.4 shows the share of deviators from the default among the civil servants who have experienced the previous payment system in their current job for at least 5 years before the introduction of the treatment. In the first year of treatment in both treatment group, the share of experienced civil servants who deviated from the default is between 55% and 60%. This includes everyone who opted for any form of payment that is not strictly monetary via the default timing of pay, including earlier monetary payouts (in the December default treatment), later monetary payouts (in the Monthly default treatment), or buying additional leave, bikes, gym memberships or other non-monetary forms of payments. In the Monthly default treatment, the share of deviators decreases substantially over the years, down to 40% for civil servants in their seventh year of treatment. At the same time, the share of deviators in the December default treatment first slightly increases up to 64% in the second year of treatment and then decreases down to 54.7% in the fifth year of treatment. [7] This suggests that if the default is to receive

[7] As the December default treatment was introduced later than the Monthly default treatment, the years of observation are less.

the payment in monetary form towards the end of the year, still, almost half of those civil servants stick to this default, and the share of civil servants who deviates from the default decreases over time.

The green line shows the share of deviators from the default among the civil servants who entered their job after the introduction of the ICB policy and are a maximum of 30 years old when starting the job of interest, so they could not experience this previous system at a different employer for a longer period of time. [8] For this group, we see a revealed preference to get money earlier than the group of experienced civil servants. This manifests in a higher deviation rate in the December default treatment and a lower deviation rate in the Monthly default treatment. The overall trends are similar, in the sense that the percentage of deviation appears to go down for both treatments.

While figure 7.4 shows the results on aggregate, table 7.3 presents the results of an OLS-regression on the outcome variable whether a civil servant deviated from the default in the respective year or not. It includes two regression specifications. The first specification is without control variables and without years 2020 and later, thus excluding any potential COVID-19 effects. The second includes control variables on household wealth, children in the household, gender and age.

| | (1)<br>Deviate=1 | (2)<br>Deviate=1 |
|---|---|---|
| December Treat | $0.256^{***}$ | $0.262^{***}$ |
| With pre-experience | 0.0127 | $0.131^{***}$ |
| December Treat × With pre-experience | $-0.105^{***}$ | $-0.122^{***}$ |
| Year of Treatment | $0.0106^{***}$ | $0.0154^{***}$ |
| CONTROLS | NO | YES |
| Constant | $0.448^{***}$ | $0.615^{***}$ |
| Observations | 403,694 | 403,694 |
| Adjusted $R^2$ | 0.007 | 0.036 |

$^{*}$ $p < 0.05$, $^{**}$ $p < 0.01$, $^{***}$ $p < 0.001$

Table 7.3.: Regression deviation across treatments and experience level

The December treatment is associated with about 26*pp* higher deviation rates across both regression specifications, while people with previous experience deviate around 13*pp* less based on the second specification.

### 7.4.2. Especially people in the December treatment with pre-experience opt back to the pre-system

After analysing the binary decision whether a civil servant deviated or adhered to the default, we now refine the analysis to the actual outcome of their timing of payment. Here, we analyse four different outcomes as introduced in section 7.3.3: *December payment* (receiving the entire ICB in December), *Monthly payment* (receiving the ICP as early as

[8] Note that in the Monthly default treatment group, there are not enough observations of new civil servants that have at least four years of treatment.

it is earned), *Imitate pre-system* (scheduling the payments as in the previous system) and *Mixed payment* (mixtures between these options and non-monetary payments).

Figure 7.5 shows these four different outcomes for all the civil servants with pre-system experience. 21.7% of the civil servants in the monthly default treatment still actively choose to get the December payment. 12% start with imitating the pre-system in the first year, a number which declines to just 5.3% percent in year seven of treatment. 20.7% choose a mixed payment outcome. The share of civil servants that sticks to the default and chooses a monthly payment schedule starts at 45.6% and increases to 58.9% in year seven of treatment.

The experienced civil servants in the December treatment show vastly different behaviour. 41.4% of them stick to the December payment, and only 10.5% choose the most immediate monthly payment outcome. Interestingly, 22% of experienced civil servants in the December default treatment choose to imitate the previous system. This is more than double compared to the 10% in the monthly default treatment. 26.1% choose a mixed payment timing.

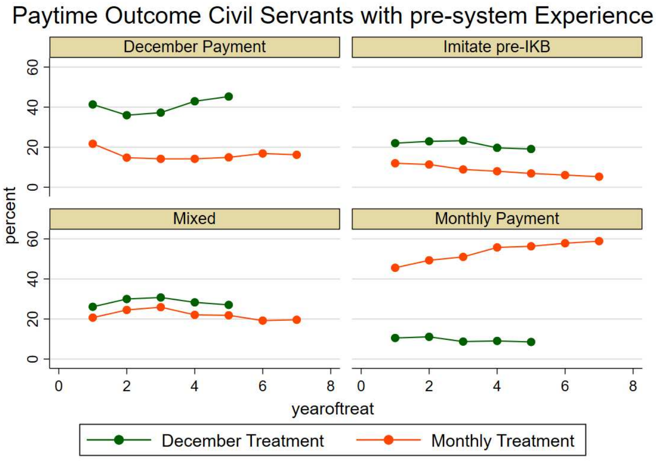


Figure 7.5.: Payout pattern choices across treatments

For the new civil servants we observe similar patterns, as can be seen in figure 7.6. Here, less people imitate the previous system, which can be explained by the fact that they have less experience with it. Yet, it shows that this system is so prominent that people still choose in many cases. Overall, new civil servants are more likely to choose the payment option that yields an earlier payout, especially by sticking to or choosing for a monthly payment.

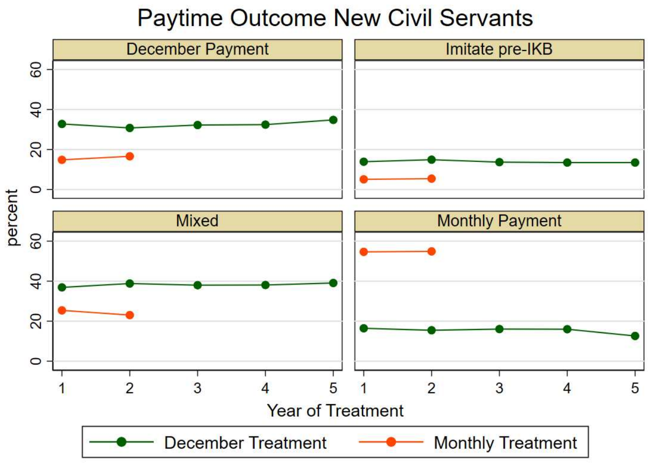


Figure 7.6.: Deviation across treatments

Table 7.4 presents the OLS regression results that focus on imitating the previous system as dependent variable. Throughout all regression specifications, civil servants in the December default treatment that have experience with the previous system are about 5% points more likely to actively choose the previous system as payment timing schedule.

| | (1) | (2) |
|---|---|---|
| | imitate pre-system=1 | |
| December Treatment | 0.0910*** | 0.0391* |
| With pre-experience | 0.0375* | 0.0254 |
| Dec Treat × With pre-experience | 0.0529*** | 0.0552*** |
| CONTROLS | NO | YES |
| ≥2020 (COVID-19) | NO | NO |
| Constant | 0.0604*** | 0.0842** |
| Observations | 403,694 | 403,694 |
| Adjusted $R^2$ | 0.008 | 0.019 |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 7.4.: Imitating the previous payment system

## 7.5. Conclusion

Traditionally, many employers in the Netherlands helped their employees create a financial buffer for the summer vacation or the holidays at the end of the year, by providing them with suitably timed double paychecks. Recently, these payout patterns have been made more flexible for many employees, to give them more agency over their salary and accommodate their specific consumption choices over time. The traditional payout pattern for Dutch civil servants was replaced by an Individual Choice Budget, with differing default payout patterns for different layers of government at different points in time. This paper analyses the results of this change in payout policies.

We find that under the more flexible salary payout regime, many people still actively decide to get their extra paycheck later than they could. This means that they effectively use their employer as a savings account. Employees who by default receive the flexible part of their salary entirely in December (the December default) are more likely to try and mimic the previous payout pattern - with a double paycheck before the summer and another one at the end of the year - than employees who by default get the flexible part of their salary payed out each month (the Monthly default). Overall, the monthly default shows the lowest amount of deviation.

People with experience of the previous system make different payout choices than newcomers to the system. Newcomers are much less likely to try to reconstruct the previous system and more frequently decide for earlier payments. However, both groups seem to deviate less from the default payout pattern over time. This is true for both the December default and the Montly default.

In our analyses, we were unable to include (detailed) information on spending. It would be interesting to focus further research on investigating to what extend the different defaults influence spending behavior and the financial health of the household.

# 8. Outlook

This dissertation presents six experimental studies aimed at understanding human behaviour through experiments. The first part includes three studies on topics of identity & belonging. Each study uses a slightly different approach to operationalise "identity" in its experimental design and modelling approach. Chapter 2 is designed for a specific case study in rural Laos. Chapter 3 builds groups on natural and artificial identity markers to build a hierarchy of groups from in-, middle- to out-group. Chapter 4 uses U.S. party affiliation to create social preferences.

Part 2 presents three studies where key elements of the experimental design are beyond the immediate control of the experimenter. Chapter 5 uses externalities that lie outside the direct incentive scheme for participants to induce differences in behaviour. Chapter 6 has an algorithm at its core whose behaviour is outside of the experimenter's control. Finally, chapter 7 is a policy analysis of a natural experiment where the researchers have not been involved in the construction of the 'experiment' at all, but can only evaluate it.

As an outlook on further research, several pathways are possible to advance the understanding of human behaviour and *how* experiments can be helpful in this endeavour.

## 8.1. Deepening the Understanding of One Phenomenon or Outcome Variable

**Meta Studies**

Meta studies are a fruitful way to condense different approaches to one research theme, such as in-group favouritism. For example, Balliet, Wu, and De Dreu (2014) find "a small to medium effect size indicating that people are more cooperative with in-group, compared to out-group, members." Also, chapter 2 finds additional token distribution to the own kin. Chapter 3 finds consistent in-group favouritism in a rule-breaking task. Chapter 4 finds in-group favouritism in a generosity task, but not in a stylised performance review task. Based on chapter 4, a meta-study on the effectiveness and elicitation of social preferences on Prolific is warranted.

**Mega Studies**

Mega studies have the comparable goal of investigating one phenomenon through multiple designs. But while a meta-study is conducted ex-post, a mega-study is planned ex-ante. This can include policy-relevant research such as the mega study by Milkman et al. (2021) on getting vaccinated. In essence, mega studies try to bring in as many treatments arms to benchmark their effect on one outcome variable. At a very small scale, chapter 6 also compares different types of interventions and their effect on delegating to an algorithm, while not attempting to analyse one of the interventions in great detail.

### Replication and Reproducibility Studies

To confirm and solidify insights of behavioural and experimental research, replication and reproducibility studies are important to uphold a high standard within science; An effort I was able to contribute to as a member of the *Management Science Reproducibility Collaboration* in Fišar et al. (2024).

## 8.2. Using Artificial Intelligence as Factor in Experimental Research

### AI and Large Language Models (LLMs) as Feature in Experiments

In business, many customer interactions on websites or even on phone lines are powered by Artificial Intelligence and LLMs, with the aim of speeding up problem-solving and improving customer satisfaction. Chapter 6 uses a comparable outset with an algorithm to aid decision-making in a selection task. Current-day extensions are to place an LLM as the feature into the experiment to analyse human reaction to, collaborate with or trust in this LLM. A most recent example is presented by Costello, Pennycook, and Rand (2024) who use LLMs to interact with participants who hold conspiracy beliefs and have the LLMs present compelling counterevidence.

### LLMs as Participants in Experiments

Furthermore, an experiment can be simulated using only LLM agents. Horton (2023) terms these agents “homo silicus” as small computational models of human agents. It is important to keep in mind that these insights are always just a reflection of the human behaviour so far that has found its way into the training data of the LLM. Therefore, inherently new behaviours in, for example, new circumstances (e.g., climate crisis, World War III, ...) can not be reliably measured through an LLM and then extrapolated on a human population. Still, “LLMs could allow researchers to pilot studies via simulation first, searching for novel social science insights to test in the real world” (Horton, 2023).

It could also be interesting to determine when and where exactly differences between LLM-agent and human-agent behaviour occur, to infer which aspects of current human behaviour have not found their way into the LLMs’ training data (yet).

### LLMs as Evaluators of Experimental Designs and Results

To complement meta-studies on a more procedural level, thousands of experimental designs and results could be fed into an LLM. This could then analyse all these experimental designs to detect trends, common patterns and unusual design approaches. This could spark a great methodological reflection of how researchers design their experiments. A comparable effort on one specific outcome variable has been done by Bhatia (2024) on exploring risky behaviour.

## 8.3. Interdisciplinary Perspectives on Experiments

### Complex Networks & Experiments

Experiments often aim to simplify a context to causally understand one mechanism. By design, complexity is difficult to include in a controllable environment. However, many human networks can be best described as complex. Among other disciplines, sociology, anthropology, physics, and the interdisciplinary field of complexity sciences aim to understand such complex networks. Combining the causal identification through experiments with complex networks could yield insights on the "systemic-frame" instead of only an "individual-frame" as described by Chater and Loewenstein (2023).

### Understanding the Moral Horizon of Experiments Better

Chapter 2 describes academic experiments as "rather exceptional kind of communicative events". To get a better understanding of the epistemological insights we can derive from such experiments, more collaboration with sociocultural anthropology is needed. This interdisciplinary reflection should not only cover how participants think and model the experiences of an experiment but also how the experimenter thinks about, models and analyses the behaviours and experiences of an experiment.

There are several pathways to continue research based on the themes in this dissertation. More pathways are possible and will most definitely present themselves in the future in a currently unknown form or shape.

# References

Abbink, K., & Harris, D. (2019). In-group favouritism and out-group discrimination in naturally occurring groups. *PloS one*, *14*(9), e0221616.

Abeler, J., & Marklein, F. (2017). Fungibility, labels, and consumption. *Journal of the European Economic Association*, *15*(1), 99–127.

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining* (pp. 2623–2631).

Alary, P. (2021). The development of commercial transactions in a" non-commercial" environment: History of a social adaptation process observed in northern laos at the end of the 20th century. *Moussons. Recherche en sciences humaines sur l'Asie du Sud-Est*(37).

Allport, G. W., Clark, K., & Pettigrew, T. (1954). The nature of prejudice.

Altmann, S., Falk, A., Heidhues, P., Jayaraman, R., & Teirlinck, M. (2019). Defaults and donations: Evidence from a field experiment. *Review of Economics and Statistics*, *101*(5), 808–826.

Alves, H., Koch, A., & Unkelbach, C. (2017). The "common good" phenomenon: Why similarities are positive and differences are negative. *Journal of Experimental Psychology: General*, *146*(4), 512.

Alves, H., Koch, A., & Unkelbach, C. (2018). A cognitive-ecological explanation of intergroup biases. *Psychological Science*, *29*(7), 1126–1133.

Amichai-Hamburger, Y. (2005). Internet minimal group paradigm. *CyberPsychology & Behavior*, *8*(2), 140–142.

Amnuayvit, T. (2017). Migration and the ethnic division of labour in siam's teak business, 1880s–1910s'. *Dreams of Prosperity: Inequality and Integration in Southeast Asia*, 131–166.

Attanasi, G., Hopfensitz, A., Lorini, E., & Moisan, F. (2016). Social connectedness improves co-ordination on individually costly, efficient outcomes. *European Economic Review*, *90*, 86–106.

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, *47*, 235–256.

Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and individual differences*, *124*, 150–159.

Balakrishnan, V., Khan, S., Fernandez, T., & Arabnia, H. R. (2019). Cyberbullying detection on twitter using big five and dark triad features. *Personality and individual differences*, *141*, 252–257.

Balliet, D., Tybur, J. M., Wu, J., Antonellis, C., & Lange, P. A. M. V. (2016). Political ideology, trust, and cooperation. *Journal of Conflict Resolution*, *62*, 797-818. doi: 10.1177/0022002716658694

Balliet, D., Wu, J., & De Dreu, C. K. (2014). Ingroup favoritism in cooperation: a meta-analysis. *Psychological bulletin*, *140*(6), 1556.

Barto, A. G. (1997). Reinforcement learning. In *Neural systems for control* (pp. 7–30). Elsevier.

Bateson, G. (1972). Steps to an ecology of mind. *New York: Ballantine*.

Bénabou, R., & Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, *30*(3), 141–64.

Ben-David, D., & Sade, O. (2021). Robo-advisor adoption, willingness to pay, and trust—before and at the outbreak of the covid-19 pandemic. *Willingness to Pay, and Trust—Before and at the Outbreak of the COVID-19 Pandemic (May 1, 2001).*

Bereby-Meyer, Y., Hayakawa, S., Shalvi, S., Corey, J. D., Costa, A., & Keysar, B. (2020). Honesty speaks a second language. *Topics in cognitive science*, *12*(2), 632–643.

Berger, B., Adam, M., Rühr, A., & Benlian, A. (2021). Watch me improve—algorithm aversion and demonstrating the ability to learn. *Business & Information Systems Engineering*, *63*(1), 55–68.

Bernhard, H., Fischbacher, U., & Fehr, E. (2006). Parochial altruism in humans. *Nature*, *442*(7105), 912.

Berrar, D. (2019). Cross-validation. In S. Ranganathan, M. Gribskov, K. Nakai, & C. Schönbach (Eds.), *Encyclopedia of bioinformatics and computational biology* (p. 542-545). Oxford: Academic Press. doi: https://doi.org/10.1016/B978-0-12-809633-8.20349-X

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly journal of economics*, *119*(1), 249–275.

Bhatia, S. (2024). Exploring variability in risk taking with large language models. *Journal of Experimental Psychology: General*.

Bogert, E., Schecter, A., & Watson, R. T. (2021). Humans rely more on algorithms than social influence as a task becomes more difficult. *Scientific reports*, *11*(1), 1–9.

Bouté, V. (2009). Names and territoriality among the phunoy. *Inter-Ethnic Dynamics in Asia: Considering the Other Through Ethnonyms, Territories and Rituals*, 79.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, *30*(7), 1145–1159.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate? *Journal of social issues*, *55*(3), 429–444.

Brune, L., Chyn, E., & Kerwin, J. (2021, July). Pay me later: Savings constraints and the demand for deferred payments. *American Economic Review*, *111*(7), 2179-2212. Retrieved from `https://www.aeaweb.org/articles?id=10.1257/aer.20191657` doi:

10.1257/aer.20191657

Bucklin, R., Lehmann, D., & Little, J. (1998). From decision support to decision automation: A 2020 vision. *Marketing Letters*, *9*, 235–246.

Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *33*(2), 220–239.

Camerer, C., & Hua Ho, T. (1999). Experience-weighted attraction learning in normal form games. *Econometrica*, *67*(4), 827–874.

Camerer, C. F. (2018). Artificial intelligence and behavioral economics. In *The economics of artificial intelligence: An agenda* (pp. 587–608). Chicago, IL 60637 U.S.A.: University of Chicago Press.

Carrier, J. G. (1997). *Meanings of the market: the free market in western culture.* Routledge.

Castelluccia, C., & Le Métayer, D. (2019). Understanding algorithmic decision-making: Opportunities and challenges.

Chater, N., & Loewenstein, G. (2023). The i-frame and the s-frame: How focusing on individual-level solutions has led behavioral public policy astray. *Behavioral and Brain Sciences*, *46*, e147.

Chen, D. L., Schonger, M., & Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88–97.

Chen, H., Harinen, T., Lee, J.-Y., Yung, M., & Zhao, Z. (2020). Causalml: Python package for causal machine learning. *arXiv preprint arXiv:2002.11631*.

Chen, Y., & Li, S. X. (2009). Group identity and social preferences. *American Economic Review*, *99*(1), 431–57.

Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., & Cagan, J. (2022). Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of ai advice. *Computers in Human Behavior*, *127*, 107018.

Chugunova, M., & 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.

Ciccarone, G., Di Bartolomeo, G., & Papa, S. (2020). The rationale of in-group favoritism: An experimental test of three explanations. *Games and Economic Behavior*, *124*, 554–568.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2001). *Introduction to algorithms* (2nd ed.). Cambridge, Massachusetts: The MIT Press.

Costa Jr, P. T., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: robust and surprising findings. *Journal of personality and social psychology*, *81*(2), 322.

Costello, T. H., Pennycook, G., & Rand, D. G. (2024, Apr). *Durably reducing conspiracy beliefs through dialogues with ai.* PsyArXiv. Retrieved from `osf.io/preprints/psyarxiv/xcwdn` doi: 10.31234/osf.io/xcwdn

De Bruyn, A., Viswanathan, V., Beh, Y. S., Brock, J. K.-U., & von Wangenheim, F.

(2020). Artificial intelligence and marketing: Pitfalls and opportunities. *Journal of Interactive Marketing*, *51*, 91–105.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114.

Dimant, E. (2024). Hate trumps love: The impact of political polarization on social preferences. *Management Science*, *70*(1), 1–31.

Downen, T., Kim, S., & Lee, L. (2024). Algorithm aversion, emotions, and investor reaction: Does disclosing the use of ai influence investment decisions? *International Journal of Accounting Information Systems*, *52*, 100664.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American statistical association*, *56*(293), 52–64.

Eil, D., & Rao, J. M. (2011). The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, *3*(2), 114–38.

Enholm, I. M., Papagiannidis, E., Mikalef, P., & Krogstie, J. (2021). Artificial intelligence and business value: A literature review. *Information Systems Frontiers*, 1–26.

Erev, I., & Roth, A. E. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American economic review*, 848–881.

Evans, G. (1995). Lao peasants under socialism and post-socialism.

Evrard, O. (2007). Interethnic systems and localized identities: the khmu subgroups (tmoy) in north-west laos. In *Social dynamics in the highlands of southeast asia* (pp. 127–159). Brill.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, *27*(8), 861–874.

Fenneman, A., Sickmann, J., Pitz, T., & Sanfey, A. G. (2021). Two distinct and separable processes underlie individual differences in algorithm adherence: Differences in predictions and differences in trust thresholds. *Plos one*, *16*(2), e0247084.

Filiz, I., Judek, J. R., Lorenz, M., & Spiwoks, M. (2021). The tragedy of algorithm aversion.

Filiz, I., Judek, J. R., Lorenz, M., & Spiwoks, M. (2022). Algorithm aversion as an obstacle in the establishment of robo advisors. *Journal of Risk and Financial Management*, *15*(8), 353.

Fišar, M., Greiner, B., Huber, C., Katok, E., Ozkes, A. I., & Collaboration, M. S. R. (2024). Reproducibility in management science. *Management Science*, *70*(3), 1343–1356.

Franco, F. M., & Maass, A. (1996). Implicit versus explicit strategies of out-group discrimination: The role of intentional control in biased language use and reward allocation. *Journal of Language and Social Psychology*, *15*(3), 335–359.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Fu, F., Tarnita, C. E., Christakis, N. A., Wang, L., Rand, D. G., & Nowak, M. A. (2012).

Evolution of in-group favoritism. *Scientific reports*, *2*(1), 460.

Galizzi, M. M., & Navarro-Martinez, D. (2019). On the external validity of social preference games: a systematic lab-field study. *Management Science*, *65*(3), 976–1002.

Gaudeul, A., & Giannetti, C. (2023). Trade-offs in the design of financial algorithms. *Available at SSRN 4432707*.

Gazit, L., Arazy, O., & Hertz, U. (2023). Choosing between human and algorithmic advisors: The role of responsibility sharing. *Computers in Human Behavior: Artificial Humans*, *1*(2), 100009.

Germann, M., & Merkle, C. (2023). Algorithm aversion in delegated investing. *Journal of Business Economics*, *93*(9), 1691–1727.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 ieee 5th international conference on data science and advanced analytics (dsaa)* (pp. 80–89).

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660.

Goerg, S. J., Hennig-Schmidt, H., Walkowitz, G., & Winter, E. (2016). In wrong anticipation-miscalibrated beliefs between germans, israelis, and palestinians. *PLoS One*, *11*(6), e0156998.

Goette, L., Han, H.-J., & Leung, B. T. K. (2020). Information overload and confirmation bias.

Goette, L., Huffman, D., & Meier, S. (2012). The impact of social ties on group interactions: Evidence from minimal groups and randomly assigned real groups. *American Economic Journal: Microeconomics*, *4*(1), 101–15.

Goldbach, C., Kayar, D., Pitz, T., & Sickmann, J. (2019). Transferring decisions to an algorithm: A simple route choice experiment. *Transportation research part F: traffic psychology and behaviour*, *65*, 402–417.

Goldbach, C., Sickmann, J., Pitz, T., & Zimasa, T. (2022). Towards autonomous public transportation: attitudes and intentions of the local population. *Transportation Research Interdisciplinary Perspectives*, *13*, 100504.

Goldberg, L. R. (1990). An alternative" description of personality": the big-five factor structure. *Journal of personality and social psychology*, *59*(6), 1216.

Goodman, B., & Flaxman, S. (2017). European union regulations on algorithmic decision-making and a “right to explanation”. *AI magazine*, *38*(3), 50–57.

Gottlieb, J., & Oudeyer, P.-Y. (2018). Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, *19*(12), 758–770.

Greene, J. D., & Paxton, J. M. (2009). Patterns of neural activity associated with honest and dishonest moral decisions. *Proceedings of the National Academy of Sciences*, *106*(30), 12506–12511.

Guelman, L., Guillén, M., & Pérez-Marín, A. M. (2012). Random forests for uplift modeling: an insurance customer retention case. In *International conference on modeling and simulation in engineering, economics and management* (pp. 123–133).

Guelman, L., Guillén, M., & Pérez-Marín, A. M. (2015). Uplift random forests. *Cybernetics and Systems*, *46*(3-4), 230–248.

Gunning, D. (2017). Explainable artificial intelligence (xai). *Defense advanced research projects agency (DARPA), nd Web*, *2*(2), 1.

Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, *18*, 43–49.

Gutierrez, P., & Gérardy, J.-Y. (2017). Causal inference and uplift modelling: A review of the literature. In *International conference on predictive applications and apis* (pp. 1–13).

Hagenbach, J., & Koessler, F. (2022). Selective memory of a psychological agent. *European Economic Review*, 104012.

Harris, D., Herrmann, B., Kontoleon, A., & Newton, J. (2015). Is it a norm to favour your own group? *Experimental Economics*, *18*, 491–521.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2). New York: Springer.

Heidhues, P., & Strack, P. (2019). Identifying present-bias from the timing of choices. *Available at SSRN*.

Heisenberg, W. (1969). *Der teil und das ganze: Gespräche im umkreis der atomphysik*. Munich: R. Piper & Co. Verlag.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2004). Overview and synthesis. *Foundations of human sociality*, 8–54.

Herm, L.-V., Heinrich, K., Wanner, J., & Janiesch, C. (2022). Stop ordering machine learning algorithms by their explainability! a user-centered investigation of performance and explainability. *International Journal of Information Management*, 102538.

Heßler, P. O., Pfeiffer, J., & Hafenbrädl, S. (2022). When self-humanization leads to algorithm aversion: what users want from decision support systems on prosocial microlending platforms. *Business & Information Systems Engineering*, *64*(3), 275–292.

High, H. (2014). *Fields of desire: Poverty and policy in laos*. NUS Press.

High, H., & Petit, P. (2013). *Introduction: The study of the state in laos* (Vol. 37) (No. 4). Taylor & Francis.

Hoelzemann, J., & Klein, N. (2021). Bandits in the lab. *Quantitative Economics*, *12*(3), 1021–1051.

Horton, J. J. (2023). *Large language models as simulated economic agents: What can we learn from homo silicus?* (Tech. Rep.). National Bureau of Economic Research.

Hruschka, D., Efferson, C., Jiang, T., Falletta-Cowden, A., Sigurdsson, S., McNamara, R., … Henrich, J. (2014). Impartial institutions, pathogen stress and the expanding social network. *Human Nature*, *25*(4), 567–579.

Huijsmans, R., & Piti, M. (2020). Rural schooling and good life in late socialist laos: Articulations, sketches and moments of 'good time'. *European Journal of East Asian Studies*, *20*(1), 163–191.

Hunt, L. T., Rutledge, R. B., Malalasekera, W. N., Kennerley, S. W., & Dolan, R. J. (2016). Approach-induced biases in human information sampling. *PLoS biology*, *14*(11), e2000638.

Imada, H. (2019). In-group favouritism in multiple social category contexts: extending generosity towards out-group members. *University of Kent (United Kingdom)*.

Ireson-Doolittle, C. (2004). *The lao: Gender, power, and livelihood*. Routledge.

Izikowitz, K. G. (1979). *Lamet: Hill peasants in french indochina* (Vol. 17). Ams Press.

Jago, A. S., & Laurin, K. (2022). Assumptions about algorithms' capacity for discrimination. *Personality and Social Psychology Bulletin*, *48*(4), 582–595.

Janneck, M., Bayerl, P. S., & Dietel, J.-E. (2013). The minimal group paradigm in virtual teams. In *International conference on human factors in computing and informatics* (pp. 457–476).

Jauernig, J., Uhl, M., & Walkowitz, G. (2022). People prefer moral discretion to algorithms: algorithm aversion beyond intransparency. *Philosophy & Technology*, *35*(1), 1–25.

Jiang, T. (2013). Cheating in mind games: The subtlety of rules matters. *Journal of Economic Behavior & Organization*, *93*, 328–336.

Johnson, E. J., & Goldstein, D. (2003). *Do defaults save lives?* (Vol. 302) (No. 5649). American Association for the Advancement of Science.

Jones, M., & Sugden, R. (2001). Positive confirmation bias in the acquisition of information. *Theory and Decision*, *50*, 59–99.

Jung, M., & Seiter, M. (2021). Towards a better understanding on mitigating algorithm aversion in forecasting: an experimental study. *Journal of Management Control*, *32*(4), 495–516.

Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards algorithms? a comprehensive literature review on algorithm aversion.

Kafka, F. (1926). *Das schloss*. Munich: Kurt Wolff Verlag.

Kelly, C., Sharot, T., et al. (2021). Individual differences in information-seeking. *Nature communications*, *12*(1), 1–13.

Kemp, J. (1988). *Seductive mirage: The search for the village community in southeast asia.* (No. 3).

Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, 33–50.

Kohavi, R., et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, pp. 1137–1145).

König, P. D., Wurster, S., & Siewert, M. B. (2022). Consumers are willing to pay a price for explainable, but not for green ai. evidence from a choice-based conjoint analysis. *Big Data & Society*, *9*(1), 20539517211069632.

Kool, W., & Botvinick, M. (2018). Mental labour. *Nature human behaviour*, *2*(12), 899–908.

Kooreman, P. (2000). The labeling effect of a child benefit system. *American Economic Review*, *90*(3), 571–583.

Kranton, R., Pease, M., Sanders, S., & Huettel, S. (2020). Deconstructing bias in social preferences reveals groupy and not-groupy behavior. *Proceedings of the National Academy of Sciences*, *117*(35), 21185–21193.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, *47*(260), 583–621.

Ku, C.-Y. (2020). When ais say yes and i say no: On the tension between ai's decision and human's decision from the epistemological perspectives. *Információs Társadalom*, *19*(4), 61–76.

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, *112*(2), 443–478.

Leib, M., Köbis, N., Soraperra, I., Weisel, O., & Shalvi, S. (2021). Collaborative dishonesty: A meta-analytic review. *Psychological Bulletin*, *147*(12), 1241.

Lesorogol, C. (2005). Experiments and ethnography: combining methods for better understanding of behavior and change. *Current Anthropology*, *46*(1), 129–136.

Leung, B. T. K. (2020). Limited cognitive ability and selective information processing. *Games and Economic Behavior*, *120*, 345–369.

Li, S. X. (2020). Group identity, ingroup favoritism, and discrimination. *Handbook of Labor, Human Resources and Population Economics*, 1–28.

Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R news*, *2*(3), 18–22.

Lindell, K., Samuelsson, R., & Tayanin, D. (1979). Kinship and marriage in northern kammu villages: the kinship model. *Sociologus*, 60–84.

Linville, P. W., Salovey, P., & Fischer, G. W. (1989). Perceived distributions of the characteristics of in-group and outgroup members: Empirical evidence and a computer simulation. *Journal of Personality and Social Psychology*, *57*(2), 165–188.

Logg, J. M. (2017). Theory of machine: When do people rely on algorithms? *Harvard Business School working paper series# 17-086*.

Luhmann, N. (1982). *The differentiation of society.* Columbia University Press.

Luhmann, N. (1984). Soziale systeme. grundriß einer allgemeinen theorie. *Frankfurt am Main: Suhrkamp*.

Luhmann, N. (1992). What is communication? *Communication theory*, *2*(3), 251–259.

Madrian, B. C., & Shea, D. F. (2001). The power of suggestion: Inertia in 401 (k) participation and savings behavior. *The Quarterly journal of economics*, *116*(4), 1149–1187.

Mahmud, H., Islam, A. N., Ahmed, S. I., & Smolander, K. (2022). What influences algorithmic decision-making? a systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, *175*, 121390.

Mahmud, H., Islam, A. N., & Mitra, R. K. (2023). What drives managers towards algorithm aversion and how to overcome it? mitigating the impact of innovation resistance through technology readiness. *Technological Forecasting and Social Change*, *193*, 122641.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 50–60.

McCarthy, J. (2007). What is artificial intelligence.

Milkman, K. L., Patel, M. S., Gandhi, L., Graci, H. N., Gromet, D. M., Ho, H., ... others (2021). A megastudy of text-based nudges encouraging patients to get vaccinated at an upcoming doctor's appointment. *Proceedings of the National Academy of Sciences*, *118*(20).

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*, 1–38.

Naar, N. (2020). Gaming anthropology: The problem of external validity and the challenge of interpreting experimental games. *American Anthropologist*, *122*(4), 784–798.

Nibud. (2019). Vakantiegeldenquête 2019. *Nibud*.

Ornetsmüller, C., Castella, J.-C., & Verburg, P. H. (2018). A multiscale gaming approach to understand farmer's decision making in the boom of maize cultivation in laos. *Ecology and Society*, *23*(2).

Palan, S., & Schitter, C. (2018). Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.

Parra, D. (2024). Eliciting dishonesty in online experiments: The observed vs. mind cheating game. *Journal of Economic Psychology*, 102715.

Parsons, C. A., & Van Wesep, E. D. (2013). The timing of pay. *Journal of Financial Economics*, *109*(2), 373–397.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Petitet, P., Attaallah, B., Manohar, S. G., & Husain, M. (2021). The computational cost of active information sampling before decision-making under uncertainty. *Nature Human Behaviour*, *5*(7), 935–946.

Ponce, F. S. (2022a). Moving away from the margins? how a chinese hydropower project made a lao community modern and comfortable. *Extracting Development: Contested Resource Frontiers in Mainland Southeast Asia*, 143–171.

Ponce, F. S. (2022b). 'eating with the people': How a chinese hydropower project changed food experiences in a lao community. *Social Anthropology/Anthropologie Sociale*, *30*(1), 1–23.

Ponce, F. S. (2023). Hydroelectric development in "china's backyard"? modernity, market integration, and (im) mobilities in northwestern laos. *Asian Anthropology*, *22*(4), 298–302.

Powers, D. M. (2020). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

Prahl, A., & Van Swol, L. (2017). Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting*, *36*(6), 691–702.

Purzycki, B. G., Apicella, C., Atkinson, Q. D., Cohen, E., McNamara, R. A., Willard, A. K., ... Henrich, J. (2016). Moralistic gods, supernatural punishment and the expansion of human sociality. *Nature*, *530*(7590), 327–330.

Radcliffe, N., & Surry, P. (1999). Differential response analysis: Modeling true responses by isolating the effect of a single action. *Credit Scoring and Credit Control IV*.

Radcliffe, N. J., & Surry, P. D. (2011). Real-world uplift modelling with significance-based uplift trees. *White Paper TR-2011-1, Stochastic Solutions*, 1–33.

Reich, T., Kaju, A., & Maglio, S. J. (2023). How to overcome algorithm aversion: Learning from mistakes. *Journal of Consumer Psychology*, *33*(2), 285–302.

Restrepo-Plaza, L., & Fatas, E. (2022). When ingroup favoritism is not the social norm a lab-in-the-field experiment with victims and non-victims of conflict in colombia. *Journal of Economic Behavior & Organization*, *194*, 363–383.

Robbins, H. (1952). Some aspects of the sequential design of experiments.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological monographs: General and applied*, *80*(1), 1.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688.

Rusch, H. (2014). The evolutionary interplay of intergroup conflict and altruism in humans: a review of parochial altruism theory and prospects for its extension. *Proceedings of the Royal Society of London B: Biological Sciences*, *281*(1794), 20141539.

Russell, S. J. (2010). *Artificial intelligence a modern approach* (4th ed.). London: Pearson Education, Inc.

Rzepakowski, P., & Jaroszewicz, S. (2012). Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems*, *32*, 303–327.

Safra, L., Lettinga, N., Jacquet, P. O., & Chevallier, C. (2022). Variability in repeated economic games: comparing trust game decisions to other social trust measures. *Royal Society Open Science*, *9*(9), 210213.

Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2008). Why can't a man be more like a woman? sex differences in big five personality traits across 55 cultures. *Journal of personality and social psychology*, *94*(1), 168.

Seabold, S., & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In *9th python in science conference.*

Sevenig, E. (2015). Social mobility in baan had-naleng: When the valuation of communality allows for a demarcation line in a multi-ethnic village in northwest laos.

Sharan, N. N., & Romano, D. M. (2020). The effects of personality and locus of control on trust in humans versus artificial intelligence. *Heliyon*, *6*(8), e04572.

Singh, S. (2014). Religious resurgence, authoritarianism, and "ritual governance": Baci rituals, village meetings, and the developmental state in rural laos. *The Journal of Asian Studies*, *73*(4), 1059–1079.

Smith, D. R., DiTomaso, N., Farris, G. F., & Cordero, R. (2001). Favoritism, bias, and error in performance ratings of scientists and engineers: The effects of power, status, and numbers. *Sex roles*, *45*, 337–358.

Snijders, C., Conijn, R., de Fouw, E., & van Berlo, K. (2023). Humans and algorithms detecting fake news: Effects of individual and contextual confidence on trust in algorithmic advice. *International Journal of Human–Computer Interaction*, *39*(7), 1483–1494.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management*, *45*(4), 427–437.

Sołtys, M., Jaroszewicz, S., & Rzepakowski, P. (2015). Ensemble methods for uplift modeling. *Data mining and knowledge discovery*, *29*, 1531–1559.

Sprenger, G. (2006a). Bone transfers: Incomplete replacement in rmeet ritual exchange.

*Taiwan Journal of Anthropology*, *4*(1).

Sprenger, G. (2006b). *Die männer, die den geldbaum fällten: Konzepte von austausch und gesellschaft bei den rmeet von takheung, laos* (Vol. 1). LIT Verlag Münster.

Sprenger, G. (2007). From kettledrums to coins: social transformation and the flow of valuables in northern laos. In *Social dynamics in the highlands of southeast asia* (pp. 161–185). Brill.

Sprenger, G. (2017). The connectivity of ethnic displays: new codes for identity in northern laos. *Asian Ethnicity*, *18*(1), 95–116.

Sprenger, G. (2021). Staying or moving: Government compliance and agency in post-zomian laos. *European Journal of East Asian Studies*, *20*(1), 83–106.

Sprenger, G. (2022). The gift as an open question. *Values and revaluations. The Transformation and Genesis of 'Values in Things' from Archaeological and Anthropological Perspectives, ed. Hans-Peter Hahn, Anja Klöckner, and Dirk Wicke*, 123–139.

Sprenger, G. (2023). Expectations of the gift: Toward a future-oriented taxonomy of transactions: Comments on "expectations of the gift"; response to comments. *Social Analysis*, *67*(1), 70–124.

Stachl, C., Hilbert, S., Au, J.-Q., Buschek, D., De Luca, A., Bischl, B., ... Bühner, M. (2017). Personality traits predict smartphone usage. *European Journal of Personality*, *31*(6), 701–722.

Stafford, C. (2020). *Economic life in the real world: logic, emotion and ethics.* Cambridge University Press.

Stanton, K., Carpenter, R. W., Nance, M., Sturgeon, T., & Andino, M. V. (2022). A multisample demonstration of using the prolific platform for repeated assessment and psychometric substance use research. *Experimental and Clinical Psychopharmacology*, *30*, 432-443. doi: 10.1037/pha0000545

Stobbe, S. P. (2015). *Conflict resolution and peacebuilding in laos: perspective for today's world.* Routledge.

Stolz, R. (2020). By means of squirrels and eggs: Kinship and mutual recognition among the khmu yuan of northern laos. *HAU: Journal of Ethnographic Theory*, *10*(2), 548–560.

Stolz, R. (2021). Living kinship, fearing spirits: sociality among the khmu of northern laos.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction.* MIT press.

Svantesson, J.-O., Tayanin, D., Lindell, K., Lundström, H., Engstrand, L., Widén, M., & Widén, B. (2014). *Dictionary of kammu yùan language and culture.* NIAS Press Copenhagen.

Tajfel, H. (1978). Social categorization, social identity and social comparison. *Differentiation between social group*, 61–76.

Thaler, R. H. (1990). Anomalies: Saving, fungibility, and mental accounts. *Journal of economic perspectives*, *4*(1), 193–205.

Trocin, C., Mikalef, P., Papamitsiou, Z., & Conboy, K. (2021). Responsible ai for digital health: a synthesis and a research agenda. *Information Systems Frontiers*, 1–19.

Vlačić, B., Corbo, L., e Silva, S. C., & Dabić, M. (2021). The evolving role of artificial intelligence in marketing: A review and research agenda. *Journal of Business Research*, *128*, 187–203.

von Wedel, P., & Hagist, C. (2022). Physicians' preferences and willingness to pay for artificial intelligence-based assistance tools: a discrete choice experiment among german radiologists. *BMC Health Services Research*, *22*(1), 1–14.

Vorapeth, K. (2024). *Geopolitics and political economy of laos*. L'Harmattan.

Weisberg, Y. J., DeYoung, C. G., & Hirsh, J. B. (2011). Gender differences in personality across the ten aspects of the big five. *Frontiers in psychology*, 178.

Wimmer, A. (2013). *Ethnic boundary making: Institutions, power, networks*. Oxford University Press.

Yamagishi, T., Jin, N., & Kiyonari, T. (1999). Bounded generalized reciprocity: Ingroup boasting and ingroup favoritism. *Advances in group processes*, *16*(1), 161–197.

Yamagishi, T., & Mifune, N. (2016). Parochial altruism: does it explain modern human group psychology? *Current Opinion in Psychology*, *7*, 39–43.

Yamagishi, T., & Yamagishi, M. (1994). Trust and commitment in the united states and japan. *Motivation and emotion*, *18*(2), 129–166.

Zhang, Y., Chen, X., et al. (2020). Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval*, *14*(1), 1–101.